PEPTI-DRIFT: SCALABLE SAFE–ACTIVE PEPTIDE GENERATION WITHOUT INFERENCE-TIME GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Therapeutic peptides are a promising drug modality, but their generation must satisfy multiple therapeutic constraints. We introduce **BindSafe-PepBench**, a fixed-budget benchmark that jointly evaluates target binding and four major safety metrics on the same generated candidates. We reveal that peptide length is a major confounder of joint binding–safety evaluation: longer peptides tend toward stronger predicted binding but less favorable predicted safety. This creates an apparent trade-off and can bias comparisons among models with different output-length distributions. We report absolute SAFE–ACTIVE yield and exact-length-matched gains to distinguish generative improvements from output-length effects. High SAFE–ACTIVE yield remains challenging, while the strongest multi-property methods rely on costly inference-time guidance. We therefore introduce **Pepti-drift**, a one-step generation framework that incorporates attraction toward target-specific binders and repulsion from liability-associated regions, requiring a single latent refinement followed by parallel decoding without inference-time guidance. Across 88 held-out targets, Pepti-drift achieves an 18.37% predicted SAFE–ACTIVE yield while retaining positive exact-length-matched gains. The resulting gains are competitive with multi-property-guided baselines while requiring **468× lower generation cost**, enabling scalable and fair high-throughput peptide design.

1 INTRODUCTION

Peptide therapeutics act by binding to specific sites on disease-relevant proteins and modulating their function (Muttenthaler et al., 2021; Wang et al., 2022). Recent advances in generative modeling provide a powerful way of exploring the vast peptide space and proposing target-specific binders (Chen et al., 2026; Kong et al., 2024b). However, therapeutic peptide generation is not merely a binder search problem; viable candidates must also avoid major safety liabilities, including toxicity, hemolysis, allergenicity, and neurotoxicity (Wang et al., 2022; Rathore et al., 2025). Safety is central to drug development and has been a major source of clinical attrition (Kola & Landis, 2004; Arrowsmith & Miller, 2013; Kuo & Barrett, 2024; Filipova et al., 2026). Strong binding and favorable safety must be achieved simultaneously (Figure 1a): a strong binder that causes hemolysis is not a viable candidate, just as a safe but nonbinding peptide is ineffective. Yet, existing evaluation protocols lack such joint assessment: binding or structural quality is typically evaluated independently of safety properties (Kong et al., 2024b;a; Zhang et al., 2026a; Zhou et al., 2026). This fragmentation makes it difficult to evaluate progress toward therapeutically viable peptide design consistently and fairly.

To address this gap, we introduce **BindSafe-PepBench**, a fixed-budget benchmark that jointly evaluates binding and four major safety liabilities on the same generated candidates. Applying this benchmark reveals an apparent negative binding–safety association that is largely explained by peptide length: longer peptides tend to exhibit stronger predicted binding but less favorable predicted safety. This finding motivates reporting both absolute SAFE–ACTIVE yield and exact-length-matched enrichment to distinguish genuine generative improvements from output-length effects.

Achieving high SAFE–ACTIVE yield remains challenging. Recent property-aware methods such as PepTune (Tang et al., 2025) and MOG-DFM (Chen et al., 2025) steer generation toward multiple therapeutic objectives through iterative inference-time guidance. More generally, guided diffusion and flow methods have demonstrated the effectiveness of optimizing objectives during the generative trajectory, including in peptide design (Wang et al., 2025). While effective, these methods repeatedly

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.

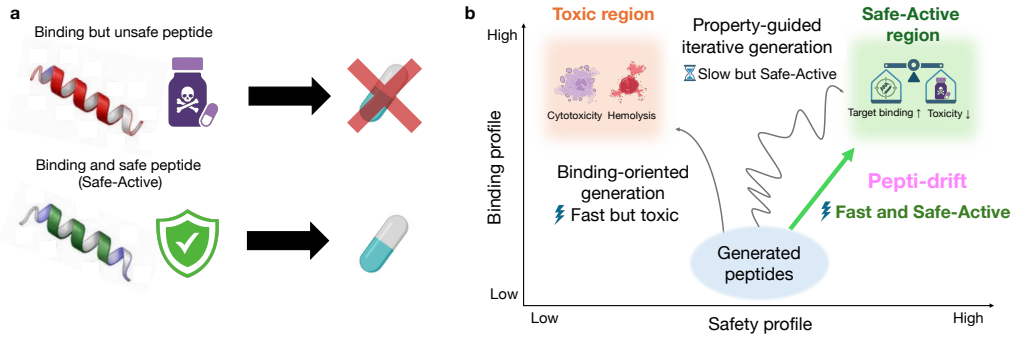


Figure 1: **Binding alone is insufficient for therapeutic peptide generation.** (a) A therapeutic candidate must satisfy both binding and safety. (b) Binding-oriented generation may overlook safety, whereas iterative property-guided methods improve joint control at additional inference cost. Pepti-drift instead amortizes multi-property control into training.

query guidance signals during sampling, incurring substantial computational cost. This overhead limits the number of candidates, targets, and design iterations that can be explored. This motivates our central research question: *How can we efficiently generate peptides that satisfy both binding and safety without costly inference-time optimization?*

To answer this question, we introduce **Pepti-drift**, a framework that amortizes multi-property control into training rather than repeatedly enforcing it during inference. Building on Drifting Models (Deng et al., 2026), Pepti-drift encodes attraction toward target-specific binders and repulsion from safety liabilities in a learned latent-space vector field, avoiding repeated optimization during sampling (Figure 1b). Given only a target protein sequence and sampled noise, Pepti-drift performs a single latent refinement followed by parallel decoding. Generation requires neither target structures nor iterative queries to binding or safety property models.

This design shifts the joint quality–cost frontier. Across 88 held-out targets, Pepti-drift achieves an 18.37% SAFE–ACTIVE yield, nearly matching the highest observed yield of 18.45%. It requires only 2.643 ms per peptide, corresponding to a $468\times$ lower generation cost than the strongest inference-guided multi-property baseline. It also retains positive exact-length-matched enrichment, indicating that its performance cannot be explained by output-length differences alone. Qualitative structural analysis further suggests that successful Pepti-drift outputs form well-buried target interfaces rather than exploiting simple one-dimensional compositional shortcuts.

These results show that joint control can be learned during training without costly inference-time search, shifting the quality–cost frontier toward faster and more practical peptide generation.

Our key contributions are threefold:

1. **Joint evaluation of binding and safety.** We introduce BINDSAFE-PEPBENCH, a fixed-budget benchmark that evaluates joint target binding and four major safety constraints on the same generated peptide.
2. **Identification of peptide length as a major confounder.** We show that peptide length largely explains the apparent binding–safety trade-off, motivating exact-length-matched evaluation to separate generative improvements from length effects.
3. **Efficient joint control without inference-time optimization.** We introduce PEPTI-DRIFT, which amortizes multi-property control into training and achieves near-best raw joint yield at $468\times$ lower generation cost than the strongest inference-guided multi-property baseline.

2 RELATED WORK

Generative modeling for biomolecular design. Generative models enable biomolecule design with desired structures and properties. Representative approaches include protein language models for sequence generation (Nijkamp et al., 2023; Wang et al., 2024b; Hayes et al., 2025), diffusion models for structure generation (Yim et al., 2023; Watson et al., 2023), and goal-directed molecular

optimization (You et al., 2018; Bengio et al., 2021; Jain et al., 2023). Our work focuses on efficient multi-property control for target-conditioned peptide generation.

Target-conditioned therapeutic peptide design. Structure-aware peptide generators such as PepGLAD (Kong et al., 2024b), PepFlow (Li et al., 2024), PPFlow (Lin et al., 2024), DiffPepBuilder (Wang et al., 2024a), PepHAR (Li et al., 2025), and PepTri (Nguyen et al., 2026) condition on target structures, binding sites, or related geometric information. PepMimic (Kong et al., 2025) additionally conditions peptide sequence–structure co-design on the binding interface of a protein binder. Beyond peptides, BindCraft (Pacesa et al., 2024; 2025) generates protein binders through iterative structure–sequence co-optimization. In contrast, PepMLM (Chen et al., 2026) and HighPepBinder (Mao et al., 2026) generate target-specific peptides directly from protein sequences. Multi-property methods such as PepTune (Tang et al., 2025) and MOG-DFM (Chen et al., 2025) further steer generation through inference-time property guidance, requiring repeated property evaluations during sampling. Related approaches to reducing iterative sampling cost include progressive distillation, consistency models, and rectified flow (Salimans & Ho, 2022; Song et al., 2023; Liu et al., 2023; Yin et al., 2024). Drifting Models (Deng et al., 2026) have also been applied to molecular generation and conformation sampling (Qiu et al., 2026; Hu, 2026). Pepti-drift extends this framework to target-conditioned peptide generation, incorporating attraction toward target-specific binders and repulsion from liability-associated regions during training to avoid iterative property queries at inference.

Benchmarks for peptide design. PepBench (Kong et al., 2024a) provides datasets and splits for protein-binding peptide design, PPBench2024 supports structure-based peptide design (Lin et al., 2024), and PepBenchmark (Zhang et al., 2026a) standardizes predictive evaluation across peptide properties. Broader biomolecular and molecular-design benchmarks emphasize multidimensional evaluation, reproducibility, oracle efficiency, or computational throughput (Ye et al., 2025; Lin et al., 2025; Brown et al., 2019; Gao et al., 2022; Liu et al., 2026). However, these protocols do not directly assess whether the same generated peptide jointly satisfies target binding and multiple safety constraints under a fixed candidate budget. BINDSAFE-PEPBENCH addresses this gap by jointly reporting SAFE–ACTIVE yield, exact-length-matched enrichment, and generation cost.

3 BINDSAFE-PEPBENCH

A useful therapeutic peptide must satisfy two requirements simultaneously: it should bind the intended protein target and avoid major safety liabilities. However, existing evaluations often assess binding and safety separately, leaving unclear whether the *same generated peptide* satisfies both. We therefore introduce BINDSAFE-PEPBENCH, a fixed-budget benchmark that evaluates joint binding and safety on individual generated candidates under a common evaluation protocol.

Fixed-budget joint generation and evaluation. Experimental validation is resource-intensive, and precise 3D target structures are often unavailable. BINDSAFE-PEPBENCH therefore considers sequence-conditioned *de novo* generation: given only a target protein sequence, each method generates a fixed set of peptides without a predefined binding site or target structure. We allocate $K = 30$ candidate slots per target, a moderate fixed budget within the range used in recent peptide-generation evaluations (Lin et al., 2024; Huang et al., 2026). We restrict candidates to 4–25 amino acids, following the peptide-length rule adopted in PepBench (Kong et al., 2024a), rather than introducing a model-specific length range. The benchmark then asks whether each candidate passes a single joint endpoint, SAFE–ACTIVE. All methods are rescored using the same external evaluators rather than their own training rewards or guidance models, avoiding direct preference for any generator's internal objectives.

For predicted binding, we first use ESMFold2 (Synthyra, 2026) to predict the 3D target–peptide complex. To avoid interpreting binding scores from low-confidence predicted complexes, we require $\mathrm{ipTM} \geq 0.5$, reflecting confidence in the relative target–peptide arrangement, and mean interface $\mathrm{pLDDT} \geq 70$, reflecting local structural confidence around the predicted interface. PRODIGY (Vangone & Bonvin, 2015; Xue et al., 2016) provides $s_{\mathrm{bind}} = -\Delta G$. For candidate-level evaluation, scores are retained only after confidence gating, and activity requires $s_{\mathrm{bind}} \geq 7$.

For predicted safety, four ESM-2-based classifiers independently evaluate toxicity, hemolysis, allergenicity, and neurotoxicity. A candidate must have a predicted liability score of at most 0.5 for all

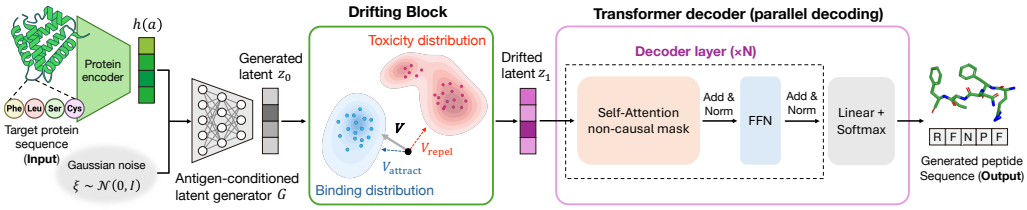


Figure 2: **Overview of the Pepti-drift framework.** A target sequence a is encoded by a protein language model, and a target-conditioned generator maps its representation and Gaussian noise to an initial peptide latent z_0. A drifting block applies a learned vector field $V = V_{\text{attract}} + V_{\text{repel}}$, combining attraction toward the target-specific binding-peptide distribution with repulsion from liability-associated peptide regions, to obtain the refined latent z_1. A non-autoregressive Transformer decoder then generates the peptide sequence from z_1 in a single parallel pass.

four properties. Finally, it must satisfy basic sequence-integrity requirements, including canonical amino acids, the allowed length range, within-target uniqueness, and exact novelty relative to the generator's training peptides. A peptide is counted as SAFE–ACTIVE only when all binding, safety, and integrity criteria are satisfied simultaneously. These outcomes are computational predictions rather than experimental measurements; complete endpoint definitions and evaluator performance are reported in Appendix B.3.

Joint yield and length-matched enrichment. Our primary metric is raw SAFE–ACTIVE yield, defined as the fraction of the fixed $K = 30$ candidate slots that satisfy the complete joint endpoint. Because every allocated slot remains in the denominator, failed or unusable generations directly reduce this yield. Raw yield can also depend on a generator's output-length distribution, because length affects both predicted binding and safety. We therefore report *Exact-length-matched enrichment@*K alongside raw yield. For each submitted candidate, we construct a random control peptide matched to the same target, submitted rank, and exact peptide length. We then compute the difference in SAFE–ACTIVE rate between the generated candidates and their matched random controls for each target and average this difference across targets. Enrichment is reported in percentage points; a positive value means that the generator produces more SAFE–ACTIVE candidates than random sequences under the same target and exact-length schedule. We additionally report SAFE–ACTIVE target coverage, defined as the fraction of targets for which at least one SAFE–ACTIVE candidate is found, together with marginal binding performance and generation cost. Further metric and control details are provided in Appendix B.

4 PEPTI-DRIFT

As motivated by BINDSAFE-PEPBENCH, therapeutic peptide generation requires joint control of binding and safety under a fixed candidate budget. Pepti-drift amortizes this multi-property control into training using the drifting paradigm (Deng et al., 2026). It learns a vector field that moves target-conditioned latent representations toward target-specific binders and away from liability-associated regions. At inference, generation requires only a single latent refinement followed by parallel non-autoregressive decoding, without iterative property evaluation or search.

One-step target-conditioned generation. Pepti-drift maps the target protein representation and sampled noise to an initial latent, followed by a single refinement step (Figure 2).

Let a denote a target protein sequence and y a reference peptide sequence. A frozen protein language model (ESM-2) produces the base target and peptide embeddings (Lin et al., 2023). The target embedding is passed through a trainable projection $h(a)$, whereas peptide references are mapped to normalized latent vectors $u_0(y) \in \mathbb{S}^{d_z-1}$ using a fixed, randomly initialized projection, where $d_z = 1024$. This fixed reference map is used exclusively to construct training targets.

An initial generator G maps the target representation and Gaussian noise $\xi \sim \mathcal{N}(0, I)$ to an initial latent z_0. A trainable drift module D then applies a single learned update, yielding the final drifted latent x (equivalently z_1),

$$z_0 = G(h(a), \xi), \qquad x = z_1 = z_0 + \eta D(z_0, h(a)), \quad \eta = 1. \tag{1}$$

At inference, the drift module D is evaluated only once during the forward pass.

Attraction–repulsion training objective. To jointly control binding and safety, we construct a dynamic training target that combines attraction toward a reference binder with repulsion from liability-associated peptides. We first ℓ_2-normalize the current generated latent x as $\bar{x} = \mathrm{norm}(x) = x/\|x\|_2$, placing it in the same normalized latent space as the peptide references. Instead of scalar rewards, we construct a spatial target $\widetilde{x}$ for the current normalized generated latent $\bar{x}$ using

$$V(\bar{x}) = \alpha_+ V_{\text{attract}}(\bar{x}) + \alpha_- V_{\text{repel}}(\bar{x}). \tag{2}$$

The *attractive component* (V_{attract}) moves the representation toward a target-matched reference binder, whereas the *repulsive component* (V_{repel}) moves it away from nearby toxicity- and hemolysis-associated references. The SAFE–ACTIVE endpoint evaluates four safety liabilities, whereas Pepti-drift constructs its repulsive field only from toxicity- and hemolysis-associated training examples. Allergenicity and neurotoxicity are kept evaluation-only, providing independent safety axes for testing generalization beyond the liabilities explicitly used during training.

We dynamically mine hard negatives whose embeddings are closest to the current latent and weight their repulsive contributions by distance. The resulting training target is $\widetilde{x} = \mathrm{sg}\left[\mathrm{norm}\left(\bar{x} + V(\bar{x})\right)\right]$, where sg denotes stop-gradient. The attraction coefficient is fixed, while the repulsion strength is gradually increased during training.

A non-causal transformer decodes the refined latent and target representations into all peptide positions in parallel, with sequence length determined by the first predicted <eos> token. Training minimizes a joint objective balancing three weighted components: a *sequence reconstruction loss* (token-level cross-entropy over non-padding positions), a *drift-alignment loss* aligning the normalized generated latent with the spatial target, and a *displacement regularizer* penalizing excessively large shifts during the one-step drift. The binding and safety evaluators used in BINDSAFE-PEPBENCH are never used as training rewards. At inference, generation requires only the target protein sequence and sampled noise. Full target construction, decoder, and optimization details are in Appendix A.

5 EXPERIMENTS

We investigate the relationship between predicted binding and safety under the fixed-budget BINDSAFE-PEPBENCH protocol and evaluate whether Pepti-drift can shift the quality–cost frontier of multi-property peptide generation through the following research questions:

(Q1) Is there a systematic tension between predicted binding and safety, and what factors drive it?
(Q2) Does Pepti-drift achieve joint success without sacrificing either the binding or safety axis?
(Q3) Can Pepti-drift match the joint SAFE–ACTIVE yield of state-of-the-art guided methods at a fraction of the computational cost?
(Q4) Does Pepti-drift find more SAFE–ACTIVE candidates per reached target than existing methods?
(Q5) Is Pepti-drift's joint success more consistent with superficial reward hacking or physically grounded 3D interactions?

We disentangle the factors driving apparent binding–safety trade-offs, revealing peptide length as a major source of evaluation bias, and demonstrate that Pepti-drift achieves a competitive SAFE–ACTIVE yield without relying on structural priors or expensive post-hoc optimization.

5.1 EXPERIMENTAL SETUP

Datasets and leakage control. The binding dataset contains 4,263 target–peptide complexes: 4,070 training, 105 validation, and 88 held-out test examples (Kong et al., 2024a). Removing exact test-peptide matches leaves 4,060 training pairs. The property dataset contains 25,480 PepBenchmark records (Zhang et al., 2026a). Toxicity, hemolysis, allergenicity, and neurotoxicity define the four safety liabilities used in the SAFE–ACTIVE endpoint. Pepti-drift's repulsion pool is constructed from toxicity- and hemolysis-labeled training peptides and is filtered against the property test set and all binding-positive peptides, leaving 3,577 sequences with no exact overlap with either protected set. Full dataset construction and leakage safeguards are described in Appendix B.2.

Baselines. We compare Pepti-drift with target-conditioned generators spanning different generation and control strategies. PepMLM (Chen et al., 2026) and High-PepBinder (Mao et al., 2026) generate

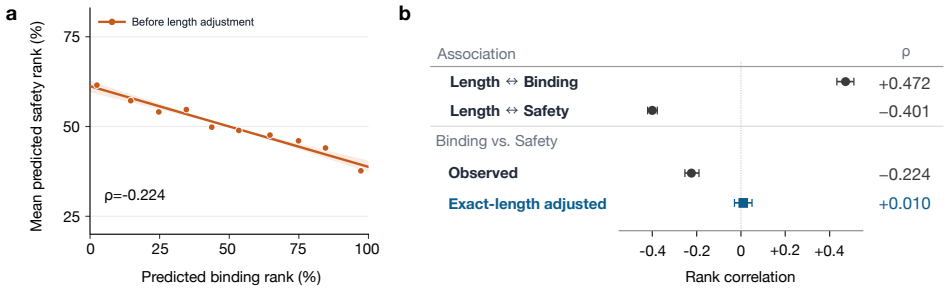


Figure 3: **Peptide length is a major driver of the apparent binding–safety tension.** (a) Predicted binding and safety exhibit a negative association before accounting for peptide length ($\rho = -0.224$). (b) Peptide length is positively associated with predicted binding and negatively associated with predicted safety. Accounting for exact peptide length nearly eliminates the observed association ($\rho = +0.010$). Error bars show target-bootstrap 95% confidence intervals.

target-specific peptides without inference-time guidance, whereas PepTune (Tang et al., 2025) and MOG-DFM (Chen et al., 2025) use multi-property guidance during inference. We additionally include a conditional autoregressive Transformer baseline trained on the same target representations and binding data as Pepti-drift. Because the primary benchmark is sequence-conditioned, methods requiring target structures or predefined binding pockets are treated as complementary rather than directly comparable baselines. Comparator configurations are provided in Appendix F.

To ensure fair evaluation, Pepti-drift and the trainable controls use benchmark-aligned splits with strict leakage safeguards. For published baselines using released checkpoints, however, exact separation between their original training data and BINDSAFE-PEPBENCH evaluation data cannot be verified.

Metrics. The primary metric is raw SAFE–ACTIVE yield at $K = 30$, interpreted alongside exact-length-matched enrichment@K, which measures improvement over target- and length-matched random controls. These metrics capture absolute joint success and improvement over a matched baseline, respectively. We also report SAFE–ACTIVE target coverage, the marginal binding-active rate, and generation cost in milliseconds per peptide. SAFE–ACTIVE target coverage is the fraction of targets with at least one SAFE–ACTIVE candidate, while binding-active rate is the fraction of submitted candidate rows satisfying the structure-confidence criteria and the binding threshold. Additional marginal binding diagnostics are reported in Appendix D.

5.2 PEPTIDE LENGTH LARGELY EXPLAINS THE APPARENT BINDING–SAFETY TRADE-OFF

To answer **(Q1)**, we examine whether predicted binding and safety exhibit a systematic tension, and what factors drive this relationship. During preliminary analysis, we observed differences in peptide-length distributions across models, suggesting that length may confound both multi-property evaluation and model comparison. To remove between-method and between-target score shifts, we rank predicted binding and the four-liability safety margin within each method–target stratum before pooling. Of 6,568 complete-case candidates, 6,356 belong to estimable strata and enter the analysis.

Before accounting for peptide length, stronger predicted binding is associated with less favorable predicted safety (Figure 3a; Spearman $\rho = -0.224$), indicating an apparent tension. Peptide length correlates oppositely with the two axes (Figure 3b): longer peptides tend to receive more favorable predicted binding scores ($\rho = +0.472$) but less favorable safety margins ($\rho = -0.401$).

This shared dependence largely explains the apparent binding–safety trade-off. When we stratify candidates by exact peptide length, restricting the analysis to method–target–length strata with sufficient observations ($n = 3{,}859$), the binding–safety association is nearly eliminated ($\rho = +0.010$). Thus, the raw negative correlation does not reflect an inherent, length-independent binding–safety antagonism. Instead, peptide length emerges as a major driver of the relationship.

This finding directly motivates the length-controlled component of BINDSAFE-PEPBENCH. Raw SAFE–ACTIVE yield remains the primary measure of absolute joint success under the fixed candidate budget, but it can be affected by a generator's output length distribution. We therefore interpret it

Table 1: **Pepti-drift remains competitive on both safety and binding axes.** Marginal performance at $K = 30$. Random uniform-length samples both peptide length and amino acids at random. Safety metrics report Safe-only and individual safety-gate pass rates. Binding metrics report binding coverage, binding-active rate, and target-level binding success. Bold and underlining mark the best and second-best among noncollapsed learned or guided methods.

	Predicted safety ↑					Predicted binding ↑		
Method	Safe-only	Tox	Hemo	All.	Neuro	Bind cov.	Bind-active	Target pass
Pepti-drift (ours)	<u>61.17</u>	<u>90.61</u>	<u>94.55</u>	80.64	<u>82.23</u>	**35.45**	<u>25.80</u>	**26.14**
PepTune	**72.73**	**96.86**	**96.14**	**88.75**	**90.83**	<u>32.54</u>	24.32	<u>25.00</u>
MOG-DFM	57.90	88.76	90.02	<u>88.42</u>	82.04	23.46	18.75	15.91
High-PepBinder	42.88	78.41	88.18	78.52	64.62	25.15	19.36	17.05
PepMLM	34.09	72.20	88.41	69.39	62.77	31.10	**27.16**	<u>25.00</u>
AR Transformer	51.93	84.13	91.48	78.18	75.49	29.13	23.37	<u>25.00</u>
Random uniform-length	35.68	71.33	70.11	84.85	57.39	20.64	16.74	9.09

jointly with Exact-length-matched enrichment@K, which asks whether a generator outperforms random peptides under the same target and exact-length schedule.

Both the unadjusted negative association and its dependence on peptide length are robust across evaluators. Across all eight binding- or safety-oracle configurations, exact-length stratification reduces the correlation to near zero ($\rho = -0.014$ to $+0.032$; all 95% CIs include zero; Appendix C.4). The unadjusted correlation is stronger for more length-dependent evaluators. These analyses answer **(Q1)**: the apparent negative association between predicted binding and safety is largely explained by peptide length rather than a length-independent antagonism, motivating exact-length-matched evaluation.

5.3 Pepti-drift remains competitive on both marginal binding and safety axes

To answer **(Q2)**, we examine whether Pepti-drift remains competitive on both the predicted binding and safety axes before applying the joint Safe–Active criterion. Table 1 summarizes the marginal quantities most directly aligned with the benchmark gates; complete safety and binding diagnostics are provided in Appendix D.

The principal generators occupy distinct marginal operating points (Table 1). PepTune is strongest on the safety axis, achieving the highest Safe-only rate (72.73%) and the highest pass rates for all four individual safety gates. PepMLM, in contrast, achieves the highest marginal binding-active rate (27.16%), but only 34.09% of its submitted candidates satisfy the binding-free Safe-only criterion.

Pepti-drift remains competitive on both axes. Its 61.17% Safe-only rate is second only to PepTune, and it ranks second in toxicity, hemolysis, and neurotoxicity pass rates. On the binding axis, Pepti-drift achieves the highest confidence-gated binding coverage (35.45%) and target pass rate (26.14%), while its marginal binding-active rate (25.80%) is second only to PepMLM.

Taken together, the marginal results address **(Q2)**: Pepti-drift's joint performance is not driven by strength on only one marginal axis. However, marginal performance alone does not establish that the two properties are satisfied by the *same peptide*. We therefore next return to the primary fixed-budget Safe–Active endpoint to evaluate joint multi-property success.

5.4 Pepti-drift shifts the joint quality–cost frontier

To answer **(Q3)**, we evaluate whether Pepti-drift achieves competitive fixed-budget Safe–Active performance without the computational overhead of inference-time guidance.

Pepti-drift achieves an 18.37% Safe–Active yield, only 0.08 percentage points below PepTune's highest raw yield of 18.45%, while requiring only 2.643 ms per peptide (Appendix D.1). This corresponds to a $468\times$ lower measured generation cost than PepTune. This result is best understood through the joint quality–cost frontier (Figure 4a). Existing methods generally trade throughput for joint candidate quality: the AR Transformer is fast but achieves lower yield, whereas PepTune achieves the highest raw yield at substantial cost. Pepti-drift shifts this frontier by combining near-best absolute joint yield, $+2.65$ positive exact-length-matched enrichment, and fast generation. Although no method dominates every metric, Pepti-drift is the only one to combine these three properties.

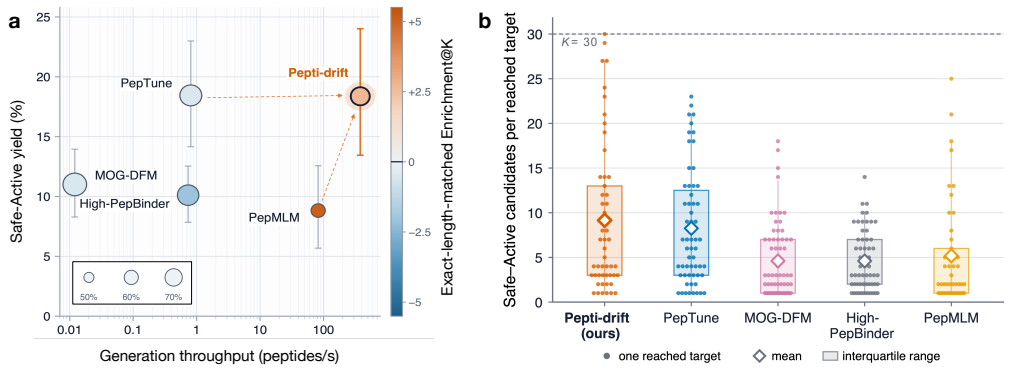


Figure 4: **Pepti-drift combines efficient joint generation with deep target-level discovery. (a)** SAFE–ACTIVE yield versus generation throughput at $K = 30$. Marker area denotes target coverage, color denotes exact-length-matched Enrichment@30, and error bars show target-bootstrap 95% confidence intervals. **(b)** Discovery depth among reached targets. Points denote individual targets, diamonds the mean, and boxes the interquartile range; the dashed line marks the $K = 30$ budget.

Compared to a speed-matched autoregressive Transformer control (2.940 ms per peptide), Pepti-drift generates candidates slightly faster (2.643 ms) while improving raw SAFE–ACTIVE yield by 3.82 percentage points, suggesting that the full Pepti-drift stack improves the performance gain. Furthermore, these efficiently generated outputs strictly satisfy the benchmark's sequence-integrity requirements: they are 100% valid and unique within targets, 99.51% exact-novel relative to training peptides, and exhibit no exact cross-target reuse (Appendix D). These findings answer **(Q3)**: Pepti-drift achieves near-best joint SAFE–ACTIVE yield at a fraction of the generation cost of inference-guided methods, without sacrificing sequence validity or novelty.

5.5 PEPTI-DRIFT YIELDS DEEPER CANDIDATE POOLS ON REACHED TARGETS

Finding one SAFE–ACTIVE candidate for a target is only a starting point: multiple distinct candidates provide more options for downstream testing and optimization. To answer **(Q4)**, we distinguish *target coverage*, the fraction of targets with at least one SAFE–ACTIVE candidate, from *candidate count per reached target*, the number of such candidates among the $K = 30$ submissions for each target.

Pepti-drift averages 9.15 SAFE–ACTIVE candidates per reached target, compared with 8.25 for PepTune and approximately 4.6–5.2 for the other published methods (Figure 4b). Thus, it provides about 11% more candidates per reached target than PepTune under the same sampling budget, at $468\times$ lower generation cost. This advantage extends to targets with many successful candidates: Pepti-drift produces at least 21 SAFE–ACTIVE candidates on seven targets and is the only compared method to produce at least 26 on any target, including all $30/30$ on one target (Appendix D.3).

These averages are conditional on each method's own reached targets and complement target coverage (Figure 4a). Together, these results answer **(Q4)**: conditional on reaching a target, Pepti-drift combines low generation cost with more candidates for follow-up on its reached targets. We next examine the predicted structures on targets where Pepti-drift provides the largest gains in candidate count.

5.6 PEPTI-DRIFT EXHIBITS PHYSICALLY GROUNDED INTERFACE FORMATION

Optimizing proxy scores raises a familiar ML concern: *reward hacking*, where models exploit evaluator imperfections to improve scores without improving the intended properties. To address **(Q5)**, we look beyond scalar scores and examine whether Pepti-drift's successful candidates exhibit plausible target interactions in their predicted structures. We quantify contact density, buried solvent-accessible surface area, and hydrophobic side-chain burial, normalizing contact counts and surface areas by peptide length (Chen et al., 2013; Munteanu et al., 2015).

We analyze 3cvl_A and 2qa9_E, selected by the prespecified rule for the largest candidate-count gains among targets reached by all five methods (Appendix D.4). Each method's representative is chosen by maximum predicted interface confidence (ipTM). On 3cvl_A, Pepti-drift's representative

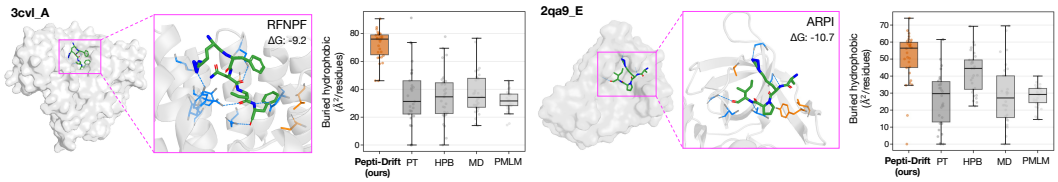


Figure 5: **Predicted interfaces and hydrophobic burial for two high-depth SAFE–ACTIVE targets.** Representative Pepti-drift complexes and enlarged target–peptide interfaces are shown for 3cvl_A (left; RFNPF) and 2qa9_E (right; ARPI), together with buried hydrophobic surface area per peptide residue across methods. Predicted ΔG is reported in kcal/mol. Peptides are shown in green; polar- and hydrophobic-contacting target residues in light blue and orange, respectively; blue dashed lines indicate putative polar contacts. Boxes show the interquartile range and median; points show individual candidates. PT, PepTune; HPB, High-PepBinder; MD, MOG-DFM; PMLM, PepMLM.

has the highest contact density and largest buried area per residue among the five representatives, with 79% of its hydrophobic side-chain surface buried against the target. Moreover, the candidate distributions show greater median buried hydrophobic surface per residue for Pepti-drift on both targets (Figure 5). On 2qa9_E, Pepti-drift's representative is at least as hydrophobic at the sequence level as MOG-DFM's, yet has a lower predicted hemolysis score (0.119 versus 0.902). Its predicted structure also buries a larger fraction of hydrophobic surface and leaves less exposed per residue. Thus, it combines favorable predicted binding and safety with substantial interface burial, without simply reducing sequence hydrophobicity. Full cross-method descriptors are reported in Appendix Table 12. Similar interface-utilization patterns are also observed on two high-depth PepTune targets, indicating that the trend extends beyond the Pepti-drift-selected targets (Appendix Figures 10 and 11).

Overall, these structural analyses answer **(Q5)**: Pepti-drift's joint success is more consistent with physically grounded 3D interface formation than with simple 1D compositional biases. The analyses are interpretive rather than independent validation and do not rule out all forms of reward hacking.

6 CONCLUSION

Before concluding, we acknowledge several limitations. This study remains entirely *in silico*: predicted binding and safety require prospective *in vitro* and *in vivo* validation to establish biological viability. Moreover, BINDSAFE-PEPBENCH captures only part of peptide developability, and future work should incorporate stability, pharmacokinetics, bioavailability, and immunogenicity.

Our component analysis also indicates that removing the drifting block produces only a modest reduction in joint performance with substantially overlapping confidence intervals in Appendix A.5. Thus, the present results support the effectiveness of Pepti-drift's full one-step generation framework, indicating that the performance gains mainly stem from the architecture as a whole rather than being dominated by the drifting block. Developing training objectives and latent-space control mechanisms that make attraction–repulsion refinement more influential is an important direction for future work.

In conclusion, we introduced BINDSAFE-PEPBENCH, revealing that the apparent negative association between predicted binding and safety is largely explained by peptide length. This finding motivated our use of exact-length-matched controls for fair joint evaluation. To this end, we proposed Pepti-drift, a one-step peptide generation framework that achieves competitive joint binding–safety control by amortizing it into the training phase, thereby avoiding costly repetitive guidance at inference time.

In our experiments, we demonstrated that shifting multi-property control from iterative inference to learned one-step refinement fundamentally improves the quality–cost trade-off of peptide generation. Specifically, Pepti-drift achieves an 18.37% SAFE–ACTIVE yield with a +2.65 percentage-point exact-length-matched enrichment. At just 2.643 ms per peptide, it incurs a 468× lower generation cost than the strongest inference-guided baseline. Overall, these results establish a more rigorous paradigm for joint evaluation and prove that competitive multi-property generation is highly scalable without costly inference-time guidance.

7 AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, language editing, and presentation of author-provided experimental results. They were not used to generate the benchmark measurements reported in this work. All AI-assisted text and interpretations were reviewed by the authors, numerical claims were checked against saved machine-readable outputs, and the cited literature was manually verified. The authors take responsibility for the final text, claims, analyses, and released artifacts.

8 ETHICS STATEMENT

This work uses public protein, peptide, and assay-derived datasets and does not involve human participants or new animal experiments. Generative peptide models are dual-use: improved design may accelerate therapeutic discovery, but could also be used to seek toxic or otherwise harmful sequences. The benchmark reports computational predictions only and must not be used to claim biological safety or to bypass experimental biosafety review. Synthesis and testing should follow institutional, legal, and assay-specific procedures. Released sequences, models, and evaluator outputs should carry provenance, licenses, intended-use constraints, and explicit warnings about predictor uncertainty.

9 REPRODUCIBILITY STATEMENT

The source code for Pepti-drift and BindSafe-PepBench will be made available upon acceptance of the paper, and data subject to non-shareable licenses are fully described in the Appendix. We provide the information required to reproduce the model, benchmark, and evaluation pipeline throughout the Appendix. Pepti-drift model definitions and training objectives are given in Appendix A, and the benchmark protocol, datasets, leakage safeguards, and evaluators are described in Appendix B. Robustness analyses and additional fixed-budget diagnostics are reported in Appendix C and D, respectively. Architecture and training hyperparameters are summarized in Appendix E, while comparator configurations, pretrained checkpoints, and generation-speed protocols are provided in Appendix F.

REFERENCES

John Arrowsmith and Philip Miller. Phase ii and phase iii attrition rates 2011–2012. *Nature Reviews Drug Discovery*, 12(8):569, 2013. doi: 10.1038/nrd4090.

Aureka AI OpenDDE Project. Folding, reasoning, and scaling with open-source drug discovery engine. *arXiv preprint arXiv:2607.03787*, 2026.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems*, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3): 1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

Jieming Chen, Nicholas Sawyer, and Lynne Regan. Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science*, 22 (4):510–515, 2013. doi: 10.1002/pro.2230.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, Jianli Wu, Audrey Stavrand, Mayumi Schaepers-Cheu, Tian Zi Wang, Divya Srijay, Connor Monticello, Pranay Vure, Rishab

Pulugurta, Sarah Pertsemlidis, Kseniia Kholina, Shrey Goel, Matthew P. DeLisa, Jen-Tsan Ashley Chi, Ray Truant, Hector C. Aguilar, and Pranam Chatterjee. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, 44:1002–1010, 2026. doi: 10.1038/s41587-025-02761-2.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. *arXiv preprint arXiv:2505.07086*, 2025.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Dilyana Filipova, Wolfgang Boomgaarden, Leyla Fox, Andrew Knight, Bettina Lickiss, Christian Maass, Gaby Neumann, Justine Watkins, Kathrin Herrmann, Lorna Ewart, et al. Need for nams: A systematic evidence synthesis revealing over half a century of drug development failure. *NAM Journal*, pp. 100119, 2026.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems*, 2022.

Marco Giulini, Victor Reys, João M. C. Teixeira, Brian Jiménez-García, Rodrigo V. Honorato, Anna Kravchenko, Xiaotong Xu, Raphaëlle Versini, Anna Engel, Stefan Verhoeven, and Alexandre M. J. J. Bonvin. HADDOCK3: A modular and versatile platform for integrative modeling of biomolecular complexes. *Journal of Chemical Information and Modeling*, 65(13):7315–7324, 2025. doi: 10.1021/acs.jcim.5c00969.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Pipi Hu. Drifting to boltzmann: Million-fold acceleration in boltzmann sampling with force-guided drifting. *arXiv preprint arXiv:2603.05527*, 2026.

Xiaoqiang Huang, Robin Pearce, and Yang Zhang. EvoEF2: Accurate and fast energy function for computational protein design. *Bioinformatics*, 36(4):1135–1142, 2020. doi: 10.1093/bioinformatics/btz740.

Yueshan Huang, Gufeng Yu, Letian Chen, Haoyang Luan, and Yang Yang. Epic: multi-objective guided diffusion for epitope design in tcr-pmhc complexes. *Bioinformatics*, 42(6):btag358, 2026. doi: 10.1093/bioinformatics/btag358.

Moksh Jain, Sharath Chandra Raparthy, Alex Hernández-García, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective GFlowNets. In *International Conference on Machine Learning*, 2023.

Ismail Kola and John Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–716, 2004. doi: 10.1038/nrd1470.

Xiangzhe Kong, Yinjun Jia, Wenbing Huang, and Yang Liu. PepBench: Dataset for protein-binding peptide design. Zenodo, 2024a.

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

Xiangzhe Kong, Rui Jiao, Haowei Lin, Ruihan Guo, Wenbing Huang, Wei-Ying Ma, Zihua Wang, Yang Liu, and Jianzhu Ma. Peptide design through binding interface mimicry with pepmimic. *Nature biomedical engineering*, pp. 1–16, 2025.

Yin-Ming Kuo and Jeffrey S Barrett. Consideration of the root causes in candidate attrition during oncology drug development. *Clin Pharmacol Drug Dev*, 13(9):952–960, 2024.

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. In *International Conference on Machine Learning*, 2024.

Jiahan Li, Tong Chen, Shitong Luo, Chaoran Cheng, Jiaqi Guan, Ruihan Guo, Sheng Wang, Ge Liu, Jian Peng, and Jianzhu Ma. Hotspot-driven peptide design via multi-fragment autoregressive extension. In *International Conference on Learning Representations*, 2025.

Haitao Lin, Odin Zhang, Huifeng Zhao, Dejun Jiang, Lirong Wu, Zicheng Liu, Yufei Huang, and Stan Z. Li. PPFLOW: Target-aware peptide design with torsional flow matching. In *International Conference on Machine Learning*, 2024.

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*, 2025.

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

Cong Liu, Milong Ren, Jiaqi Guan, Chengyue Gong, Jinyuan Sun, Xinshi Chen, and Wenzhi Xiao. Protdbench: A unified benchmark of protein binder design and evaluation, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Qingyi Mao, Silong Zhai, Sen Cao, et al. High-PepBinder: A plm-guided latent diffusion framework for affinity-aware target-specific peptide design. *Journal of Chemical Information and Modeling*, 66(8):4863–4877, 2026. doi: 10.1021/acs.jcim.6c00253.

Cristian R Munteanu, Antonio C Pimenta, Carlos Fernandez-Lozano, Andre Melo, Maria NDS Cordeiro, and Irina S Moreira. Solvent accessible surface area-based hot-spot detection methods for protein–protein and protein–nucleic acid interfaces. *Journal of chemical information and modeling*, 55(5):1077–1086, 2015.

Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. *Nature reviews Drug discovery*, 20(4):309–325, 2021.

Ngoc-Quang Nguyen, Jaeyoon Jung, Seijung Kim, Sunkyu Kim, and Jaewoo Kang. PepTri: Tri-guided all-atom diffusion for peptide design via physics, evolution, and mutual information. In *International Conference on Learning Representations*, 2026.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978.e3, 2023. doi: 10.1016/j.cels.2023.10.002.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, et al. Bindcraft: one-shot design of functional protein binders. *BioRxiv*, pp. 2024–09, 2024.

Martin Pacesa, Lennart Nickel, Christian Schellhaas, Joseph Schmidt, Ekaterina Pyatova, Lucas Kissling, Patrick Barendse, Jagrity Choudhury, Srajan Kapoor, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Adrie H. Westphal, Simon Lindhoud, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Daan C. Swarts, Alex J. Vecchio, Bernard L. Schneider, Sergey Ovchinnikov, and Bruno E. Correia. One-shot design of functional protein binders with bindcraft. *Nature*, 646(8084):483–492, 2025. doi: 10.1038/s41586-025-09429-6.

Jiangjie Qiu, Yijun Li, Wentao Li, and Xiaonan Wang. Driftingmol: Decoder-coupled drift for one-pass property-conditional molecular generation. *arXiv preprint arXiv:2605.24841*, 2026.

Anand Singh Rathore, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta, and Gajendra PS Raghava. Prediction of hemolytic peptides and their hemolytic concentration. *Communications Biology*, 8(1):176, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Synthyra. ESMFold2-Experimental-Fast-Cutoff2025 model card. `https://huggingface.co/Synthyra/ESMFold2-Experimental-Fast-Cutoff2025`, 2026.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. PepTune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *International Conference on Machine Learning*, 2025.

Anna Vangone and Alexandre M. J. J. Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, 4:e07454, 2015.

Fanhao Wang, Yuzhe Wang, Laiyi Feng, Changsheng Zhang, and Luhua Lai. Target-specific de novo peptide binder design with diffpepbuilder. *Journal of Chemical Information and Modeling*, 64(24): 9135–9149, 2024a. doi: 10.1021/acs.jcim.4c00975.

Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy*, 7(1):48, 2022.

Luran Wang, Chaoran Cheng, Yizhen Liao, Yanru Qu, and Ge Liu. Training free guided flow-matching with optimal control. In *International Conference on Learning Representations*, 2025.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, 2024b.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Li C. Xue, João P. G. L. Rodrigues, Panagiotis L. Kastritis, Alexandre M. J. J. Bonvin, and Anna Vangone. PRODIGY: A web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 2016.

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*, 2025.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. In *International Conference on Machine Learning*, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay S. Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, 2018.

Jiahui Zhang, Rouyi Wang, Kuangqi Zhou, Tianshu Xiao, Lingyan Zhu, Yaosen Min, and Yang Wang. Pepbenchmark: A standardized benchmark for peptide machine learning. *arXiv preprint arXiv:2604.10531*, 2026a.

Yinuo Zhang, Sophia Tang, Tong Chen, Elizabeth Mahood, Sophia Vincoff, and Pranam Chatterjee. Peptiverse: A unified platform for therapeutic peptide property prediction. *bioRxiv*, 2026b.

Ziheng Zhou, Huiyu Luo, Xiaohu Zhu, Nan Wang, Xuebiao Qin, Chaoyan Zhang, and Jun Yan. Ampbench-mt: A homology-controlled benchmark for antimicrobial peptide potency, spectrum, and safety prediction. *arXiv preprint arXiv:2607.25518*, 2026.

A PEPTI-DRIFT IMPLEMENTATION DETAILS

A.1 TRAINING DATA

Binding-positive examples were obtained from the official PepBench splits (Kong et al., 2024a). After removing 10 training rows whose peptide matched one of five held-out test peptides, the exported train/validation/test sets contained 4,060/105/88 protein–peptide pairs. Deduplication of identical protein-sequence–peptide-sequence tuples by the training loader resulted in 3,953/104/88 effective pairs, respectively. For negative repulsion, we used canonical 1–50-residue peptides labeled toxic and/or hemolytic in the PepBenchmark training split (hybrid split, seed 0) (Zhang et al., 2026a). Deduplication produced 3,761 sequences; removing 184 sequences overlapping the property test set and excluding all exact binding-positive overlaps left 3,577 negatives. A deterministic 90/10 split (seed 42) assigned 3,219 negatives to training and 358 to validation. The held-out test set was used neither for optimization nor checkpoint selection.

A.2 TRAINABLE MODEL AND FIXED PEPTIDE REFERENCE MAP

We distinguish the parameters optimized by Pepti-drift from the fixed components used to construct its training targets.

Let

$$\Theta = \{\theta_A, \theta_G, \theta_D, \theta_{\mathrm{dec}}\} \tag{3}$$

denote the trainable parameters of the target projection, conditional latent generator, one-step drift module, and non-autoregressive decoder, respectively.

Target protein and peptide sequences are first represented by frozen, mean-pooled ESM-2 3B embeddings,

$$e_A(a) \in \mathbb{R}^{2560}, \tag{4}$$

$$e_P(y) \in \mathbb{R}^{2560}. \tag{5}$$

The target embedding is mapped through a trainable two-layer projection,

$$h_{\theta_A}(a) = \phi_{A,\theta_A}(e_A(a)) \in \mathbb{R}^{2560}. \tag{6}$$

This projection lies on the differentiable path to the generator, drift module, and decoder and is optimized jointly with the remainder of the model.

Peptide references are instead mapped by a separate fixed two-layer MLP,

$$r_0(y) = \phi_{P,0}(e_P(y)) \in \mathbb{R}^{d_z}, \tag{7}$$

$$u_0(y) = \mathrm{norm}(r_0(y)) = \frac{r_0(y)}{\|r_0(y)\|_2} \in \mathbb{S}^{d_z-1}, \qquad d_z = 1024. \tag{8}$$

The map $\phi_{P,0}$ has architecture

$$2560 \to 1024 \to \mathrm{ReLU} \to 1024. \tag{9}$$

Its parameters are randomly initialized once under the model initialization seed and are held fixed at those initial values throughout Pepti-drift training. In particular, $\phi_{P,0}$ is not pretrained and is not optimized by any loss in this work.

The purpose of $\phi_{P,0}$ is therefore only to provide a stationary d_z-dimensional reference coordinate system in which positive and safety-associated peptide examples can be compared with generated latents. We do not interpret this projection itself as a learned binding- or safety-specific representation.

For each positive training pair (a_i, y_i^+), the corresponding attractive reference is

$$u_i^+ = u_0(y_i^+). \tag{10}$$

The same map is applied to all safety-associated peptide references,

$$u_j^- = u_0(y_j^-). \tag{11}$$

Because all reference vectors are ℓ_2-normalized,

$$\|u - v\|_2^2 = 2 - 2u^\top v, \qquad u, v \in \mathbb{S}^{d_z-1}, \tag{12}$$

so ranking normalized references by squared Euclidean distance is equivalent to ranking them by cosine similarity.

A.3 One-step pushforward and detached target construction

For each training example i, we sample

$$\xi_i \sim \mathcal{N}(0, I_{d_z}) \tag{13}$$

and compute

$$z_{i,0} = G_{\theta_G}\left(h_{\theta_A}(a_i), \xi_i\right), \tag{14}$$

$$\Delta_{i,0} = D_{\theta_D}\left(z_{i,0}, h_{\theta_A}(a_i)\right), \tag{15}$$

$$x_i = z_{i,1} = z_{i,0} + \eta\Delta_{i,0}, \qquad \eta = 1. \tag{16}$$

Equivalently,

$$x_i = g_\Theta(a_i, \xi_i), \tag{17}$$

with conditional pushforward distribution

$$q_{\Theta,a_i} = (g_\Theta(a_i, \cdot))_{\#}\mathcal{N}(0, I_{d_z}). \tag{18}$$

We normalize the refined latent,

$$\bar{x}_i = \mathrm{norm}(x_i). \tag{19}$$

Hard-negative selection is performed using a detached copy of the current latent. For each available negative reference,

$$d_{ij}^- = \frac{1}{d_z}\left\|\mathrm{sg}(\bar{x}_i) - u_{ij}^-\right\|_2^2. \tag{20}$$

Let $\mathcal{H}_i$ contain the indices of the K_{neg} smallest distances. For $j \in \mathcal{H}_i$,

$$w_{ij}^- = \frac{\exp(-d_{ij}^-/\tau)}{\sum_{\ell\in\mathcal{H}_i}\exp(-d_{i\ell}^-/\tau)}. \tag{21}$$

We expand the attraction–repulsion field for training example i at epoch e as

$$V_i(\bar{x}_i; e) = \alpha_+ V_{\text{attract}}(\bar{x}_i) + \alpha_-(e)V_{\text{repel}}(\bar{x}_i), \tag{22}$$

where $\alpha_+ = 1$. Specifically,

$$V_i(\bar{x}_i; e) = \left(u_i^+ - \bar{x}_i\right) + \alpha_-(e)\sum_{j\in\mathcal{H}_i} w_{ij}^-\left(\bar{x}_i - u_{ij}^-\right), \tag{23}$$

where the repulsion coefficient is linearly warmed up as

$$\alpha_-(e) = \alpha_-^{\max}\min\left(1, \frac{\max(0, e - e_{\mathrm{start}})}{e_{\mathrm{warmup}}}\right). \tag{24}$$

Here, $\alpha_-^{\max}$ denotes the final repulsion strength, e_{start} the epoch at which repulsion is activated, and e_{warmup} the number of epochs over which $\alpha_-(e)$ increases linearly from zero to $\alpha_-^{\max}$.

The detached alignment target is

$$\widetilde{x}_i = \mathrm{sg}\left[\mathrm{norm}\left(\bar{x}_i + V_i(\bar{x}_i; e)\right)\right]. \tag{25}$$

The stop-gradient makes target construction asymmetric: $\widetilde{x}_i$ is recomputed from the current forward-pass latent but is treated as fixed during the corresponding backward pass. Consequently, the drift loss updates the generator and drift module through $\bar{x}_i$, but does not update the positive reference map, negative reference map, hard-negative selection, or kernel weights.

A.4 NON-CAUSAL DECODER AND OPTIMIZATION OBJECTIVE

Let $L = 50$ be the maximum decoder length. For position $t \in \{1, \dots, L\}$, the initial decoder representation is

$$H_{i,t}^{(0)} = q_t + W_z x_i + W_a h_{\theta_A}(a_i), \tag{26}$$

where q_t is a learned positional query. A Transformer encoder without a causal mask processes all positions jointly,

$$H_i = \mathcal{T}_{\theta_{\text{dec}}}\left(H_i^{(0)}\right), \tag{27}$$

followed by a token projection and softmax,

$$\ell_{i,t} = W_o H_{i,t} + b_o, \tag{28}$$

$$p_{i,t} = \text{softmax}(\ell_{i,t}). \tag{29}$$

The resulting conditional factorization is

$$P_\Theta(s_{i,1:L} \mid a_i, \xi_i) = \prod_{t=1}^{L} P_\Theta\left(s_{i,t} \mid x_i, h_{\theta_A}(a_i)\right). \tag{30}$$

Although the distribution factorizes across positions, their representations interact through non-causal self-attention before token prediction. All positions are therefore decoded in one parallel pass. The generated sequence is truncated at the first <eos> token output; thus, the length is not supplied explicitly and $L = 50$ specifies only the maximum decoded length. Let

$$\Omega = \{(i,t) : s_{i,t}^+ \neq \texttt{<pad>}\} \tag{31}$$

denote the non-padding target positions. The sequence loss is

$$\mathcal{L}_{\text{seq}} = -\frac{1}{|\Omega|} \sum_{(i,t)\in\Omega} \log P_\Theta\left(s_{i,t} = s_{i,t}^+ \mid a_i, \xi_i\right). \tag{32}$$

The drift-alignment loss is

$$\mathcal{L}_{\text{drift}} = \frac{1}{Bd_z} \sum_{i=1}^{B} \|\text{norm}(x_i) - \widetilde{x}_i\|_2^2. \tag{33}$$

We have $\|\text{norm}(x_i) - \widetilde{x}_i\|_2^2 = 2 - 2\,\text{norm}(x_i)^\top \widetilde{x}_i$, and thus, minimizing $\mathcal{L}_{\text{drift}}$ is equivalent to maximizing the cosine similarity between the refined latent and the attraction–repulsion target. The displacement regularizer is

$$\mathcal{L}_{\text{reg}} = \frac{1}{Bd_z} \sum_{i=1}^{B} \|\Delta_{i,0}\|_2^2. \tag{34}$$

The complete objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seq}}\mathcal{L}_{\text{seq}} + \lambda_{\text{drift}}\mathcal{L}_{\text{drift}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{35}$$

Optimization is performed over the trainable model parameters Θ,

$$\Theta^\star = \arg\min_{\Theta} \; \mathbb{E}\left[\mathcal{L}_{\text{total}}\right]. \tag{36}$$

The ESM-2 encoders and peptide reference map are fixed:

$$\nabla_{\omega_{\text{ESM}}}\mathcal{L}_{\text{total}} = 0, \qquad \nabla_{\phi_{P,0}}\mathcal{L}_{\text{total}} = 0. \tag{37}$$

In contrast,

$$\nabla_{\theta_A}\mathcal{L}_{\text{total}}, \quad \nabla_{\theta_G}\mathcal{L}_{\text{total}}, \quad \nabla_{\theta_D}\mathcal{L}_{\text{total}}, \quad \nabla_{\theta_{\text{dec}}}\mathcal{L}_{\text{total}} \tag{38}$$

are generally nonzero.

At inference, the target peptide, safety-derived negative references, $\phi_{P,0}$, and the attraction–repulsion target are not evaluated. Generation requires only the target protein representation, sampled Gaussian noise, one generator evaluation, one learned drift update, and one parallel decoder evaluation.

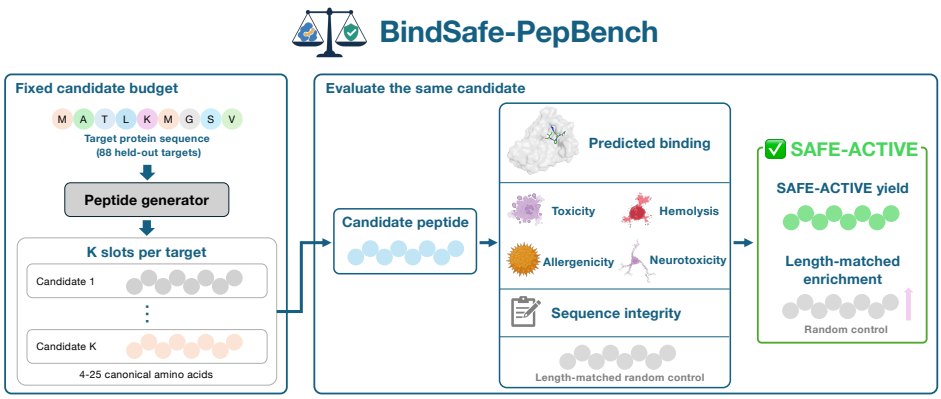


Figure 6: **Overview of BindSafe-PepBench.** Given only a target protein sequence, each method generates a fixed budget of K candidate peptides for each of 88 held-out targets. The same candidate is jointly evaluated for predicted target binding, four safety liabilities (toxicity, hemolysis, allergenicity, and neurotoxicity), and sequence integrity (validity, uniqueness, novelty, reuse). Candidates satisfying all criteria are counted as SAFE–ACTIVE. The benchmark reports raw SAFE–ACTIVE yield together with exact-length-matched enrichment against random peptides matched to the same target and peptide length.

A.5 COMPONENT ANALYSIS

We additionally evaluate several components of Pepti-drift using an independently frozen seed-45 run (Table 2). The full model achieves the highest SAFE–ACTIVE rate (17.80%) and exact-length-matched enrichment (+7.54 percentage points). Removing the drifting block results in only a modest reduction, whereas removing the one-step Transformer or replacing ESM-2 representations with one-hot features progressively reduces joint performance. Because this analysis uses a single frozen training seed and the confidence intervals overlap substantially, we interpret these results as descriptive component-level evidence rather than statistically separable causal effects of individual components.

Table 2: **Component analysis of Pepti-drift at $K = 30$.** Results are from an independently frozen seed-45 run. Enrichment is the percentage-point improvement over the target- and exact-length-matched random control; brackets show target-bootstrap 95% confidence intervals.

Condition	SA Enrichment ↑	SA rate ↑
Full	**+7.54** [3.33, 12.27]	**17.80% [12.54, 23.33]**
No drifting block	+7.05 [2.50, 11.63]	17.31% [12.12, 22.84]
No one-step Transformer	+3.90 [0.68, 7.54]	14.17% [10.00, 18.56]
No ESM-2 (one-hot)	+2.35 [−1.02, 5.91]	12.61% [8.71, 17.27]

B BINDSAFE-PEPBENCH DETAILS

B.1 TASK, ENDPOINT, AND MATCHED CONTROL

For each target protein sequence a_t, a submission provides an ordered list of up to $K = 30$ peptide candidates. The primary track restricts peptides to 4–25 canonical amino acids. Every submitted slot remains part of the fixed budget, so invalid, duplicated, out-of-range, underfilled, or unevaluable slots cannot be removed after generation (Figure 6).

A candidate is SAFE–ACTIVE only if it satisfies the common structure and binding criteria

$$\text{ipTM} \geq 0.5, \qquad \text{interface pLDDT} \geq 70, \qquad s_{\text{bind}} \geq 7,$$

and all four safety scores—toxicity, hemolysis, allergenicity, and neurotoxicity—are at most 0.5. It must additionally be canonical, have length 4–25, be an exact non-match to generator training peptides, be the first occurrence within its target, and be reused exactly across at most two targets.

Raw SAFE–ACTIVE yield is the primary metric and uses the full allocated candidate budget as its denominator. Exact-length-matched Enrichment@K is reported alongside it as a key complementary metric. The two measures are interpreted jointly to distinguish absolute joint success from improvement over a matched random baseline.

For Enrichment@K, every observed candidate is paired with a uniformly sampled random peptide that preserves its target, submitted rank, and exact peptide length. Because peptide length can affect both binding and safety evaluator outputs, this matching prevents Enrichment@K from rewarding methods merely for producing a favorable length distribution. Enrichment is computed as the target-macro difference in SAFE–ACTIVE rate between the method and its paired control. Percentile intervals use 2,000 target-bootstrap replicates with seed 0.

The matched control therefore controls for target and exact peptide length, but does not explicitly match amino-acid composition, net charge, hydrophobicity, or aromatic content. Enrichment@K should consequently be interpreted as improvement over random sequences of the same length, rather than over a fully composition-matched generator.

B.2 DATASET CONSTRUCTION AND LEAKAGE SAFEGUARDS

The binding dataset is constructed from the official PepBench split and contains 4,263 target–peptide complexes: 4,070 training, 105 validation, and 88 held-out test examples (Kong et al., 2024a). The identifiers correspond to 2,886 unique target sequences, and exact target sequences do not cross from training to validation or test. Held-out native peptides have lengths 4–21 with median 10.

For trainable generators, 10 training rows whose peptide sequence exactly matches one of five held-out test peptides are removed, leaving 4,060 binding pairs. The safety-derived repulsive pool starts from 3,761 unique positive training-split property peptides. Removing 184 sequences that occur in the property test set leaves 3,577 sequences, with zero exact overlap with the property test set or with any binding-positive peptide. A total of 162 sequences also occur in the property validation split; these validation labels are not used to evaluate generated candidates.

At the target level, the held-out split is also dissimilar from training under a simple sequence-overlap audit: the median held-out target shares 5.0% of its 4-mers with its most similar training target, while 4.5% and 2.3% of targets exceed 30% and 50% containment, respectively. For short 4–25-residue peptides, nearest-neighbor sequence identity has a high chance floor and is therefore not used as a novelty threshold. The benchmark instead applies exact sequence novelty and exact reuse constraints.

B.3 PROPERTY EVALUATOR PERFORMANCE

Table 3 reports held-out performance of the four frozen property models that define the SAFE–ACTIVE endpoint. The positive class is the unsafe class for every task. Benchmark gating uses a common score threshold of 0.5; this threshold is intended as a consistent benchmark decision rule rather than as a calibrated biological risk boundary.

Table 3: Held-out performance of the four property evaluators used in the SAFE–ACTIVE endpoint. BA is balanced accuracy and MCC is Matthews correlation coefficient.

Property	Test n	ROC-AUC	PR-AUC	BA	MCC
Toxicity	407	0.756	0.628	0.663	0.344
Hemolysis	432	0.825	0.674	0.733	0.462
Allergenicity	255	0.874	0.758	0.759	0.533
Neurotoxicity	317	0.700	0.582	0.668	0.329

C ROBUSTNESS OF BINDING–SAFETY EVALUATION

Because the central binding–safety finding depends on computational models on both axes, it is important to distinguish conclusions that persist across evaluator choices from those specific to a particular oracle. We therefore audit the observed association in three steps. We first characterize

the binding–safety association under the primary benchmark evaluators and examine whether it is driven by particular methods or targets. We then vary the binding oracle while holding safety fixed, and finally vary the safety oracle while keeping the generated candidate sets fixed. Together, these analyses provide a two-sided stress test of the computational evaluation pipeline. They are intended to assess robustness of the benchmark conclusions, not to replace experimental validation.

C.1 Primary association and method-level diagnostics

We first establish the binding–safety relationship under the primary benchmark evaluators. The descriptive analysis retains valid, in-range, nonduplicate candidates with complete binding and safety scores. Binding and safety are ranked separately within each method–target stratum before pooling, preventing between-method or between-target mean shifts from inducing the reported association. Confidence intervals use 20,000 target-cluster bootstrap replicates. Target-specific permutation tests use 20,000 permutations followed by Benjamini–Hochberg correction (Benjamini & Hochberg, 1995).

The primary safety utility uses only the four liabilities defining the SAFE–ACTIVE endpoint. Across 6,568 complete candidates from 12 submissions, the stratified binding–safety association is negative ($\rho = -0.224$, 95% CI -0.253 to -0.195), with a negative within-target association for 73 of 82 estimable targets. Toxicity shows the strongest component-level association with binding ($\rho = -0.277$).

An auxiliary utility that additionally includes nonfouling yields a weaker but still negative aggregate association ($\rho = -0.131$, 95% CI -0.160 to -0.098). We therefore focus the main analysis on the four liabilities that directly define the benchmark endpoint.

Figure 7 shows that the pooled result is not driven by a single generator. Under the four-liability utility, the association is negative for every principal learned or guided method, although its magnitude varies across methods and individual method-level estimates need not all differ from zero. Estimates range from $\rho = -0.400$ for High-PepBinder (95% CI -0.469 to -0.321) to $\rho = -0.015$ for PepMLM (-0.095 to 0.071), the only method whose interval spans zero.

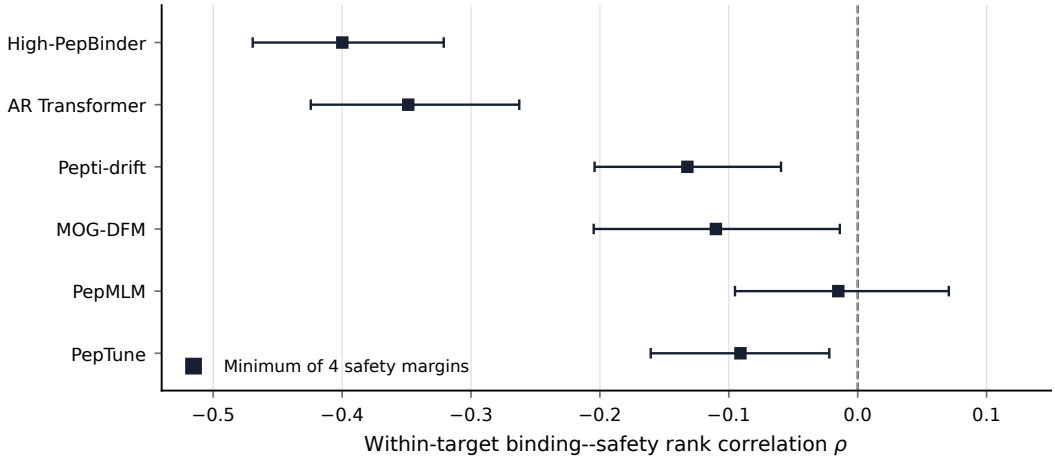


Figure 7: Method-specific within-target predicted binding–safety associations for the principal learned or guided methods. Points are Spearman estimates and bars are target-bootstrap 95% confidence intervals. The safety score is the minimum of the four liability margins defining the SAFE–ACTIVE endpoint, excluding the auxiliary nonfouling margin. The dashed line denotes zero association.

C.2 Robustness to independent binding oracles

Having established the association under the primary benchmark pipeline, we first test whether its direction depends on the binding oracle. The primary benchmark uses ESMFold2 together with PRODIGY. We therefore add two independent scoring tracks: OpenDDE+EvoEF2 and

OpenDDE+HADDOCK3 (Aureka AI OpenDDE Project, 2026; Huang et al., 2020; Giulini et al., 2025). The latter two share an OpenDDE pose but use different scoring functions, whereas the primary pipeline differs in both structure predictor and scorer.

All scores are oriented so that larger values indicate more favorable predicted binding. Because the three scoring functions have different units and calibration, we compare continuous within-method–target ranks and do not transfer the PRODIGY activity threshold across oracles. The OpenDDE panel contains five learned or guided methods and the uniform-random control over all 88 targets. OpenDDE produced 15,834/15,834 requested poses, and both EvoEF2 and HADDOCK3 returned scores for every pose; 15,596 candidates remain after the common safety and integrity join.

Table 4 reports the four-liability analysis used in the main benchmark. These values are not expected to exactly match the primary complete-case statistic above: this analysis uses continuous binding scores without the benchmark's binary structure-confidence and binding gates, and the candidate set differs across oracle panels. Despite these differences, all three target-macro confidence intervals exclude zero, preserving the negative aggregate direction under substantially different binding pipelines.

Table 4: Binding–safety association under three binding oracles. The safety utility is the minimum signed margin over the four liabilities defining the SAFE–ACTIVE endpoint. Intervals are target-bootstrap 95% confidence intervals. Cross-oracle differences in correlation magnitude are not interpreted as differences in oracle quality.

Oracle	n	Strat. ρ	Target-macro ρ [95% CI]	Negative targets
ESMFold2+PRODIGY	23,993	-0.235	-0.235 [-0.250, -0.219]	100.0%
OpenDDE+EvoEF2	15,596	-0.098	-0.098 [-0.117, -0.078]	85.2%
OpenDDE+HADDOCK3	15,596	-0.137	-0.137 [-0.154, -0.120]	94.3%

Table 5 resolves the same quantity by method. The association is negative for every method under all three oracles with one exception: PepTune, whose estimate is close to zero under ESMFold2+PRODIGY, becomes marginally positive under both OpenDDE tracks. Pepti-drift shows the most stable point estimates across the three pipelines (-0.118, -0.114, -0.104), whereas HighPepBinder and the uniform-random control show the strongest negative associations throughout. Because only point estimates are reported at this resolution, entries close to zero should not be interpreted as method-level significance tests.

Table 5: Method-level within-target binding–safety association under three binding oracles. The safety utility is the minimum signed margin over the four liabilities defining the SAFE–ACTIVE endpoint. Entries are Spearman ρ point estimates over $n = 2{,}416$–$2{,}640$ candidates per method and oracle across 88 targets. Unlike Table 4, they carry no confidence intervals and are computed without the benchmark's binary structure-confidence and binding gates.

Method	ESMFold2+PRODIGY	OpenDDE+EvoEF2	OpenDDE+HADDOCK3
Pepti-drift (ours)	-0.118	-0.114	-0.104
PepMLM	-0.057	-0.006	-0.020
PepTune	-0.084	0.034	0.034
MOG-DFM	-0.081	-0.062	-0.090
High-PepBinder	-0.411	-0.260	-0.361
Random uniform-length	-0.405	-0.191	-0.300

The three binding oracles disagree much more strongly about individual candidates than the aggregate statistics might suggest. Within method–target strata, the Spearman correlation between candidate binding scores is only 0.136 for ESMFold2+PRODIGY versus OpenDDE+EvoEF2 and 0.207 for ESMFold2+PRODIGY versus OpenDDE+HADDOCK3, rising to 0.569 between the two OpenDDE tracks that share a pose. Oracle disagreement is therefore concentrated in *which* candidate ranks highest rather than in the aggregate direction of the binding–safety association.

We consequently interpret the repeated negative direction as robust to the binding pipeline, while making no claim that the binding oracles are interchangeable or that any predicted ranking establishes

experimental affinity. This establishes robustness to substantial changes on the binding side of the evaluation; we next perform the complementary audit on the safety side.

C.3 Robustness to independent safety oracles

The preceding analysis perturbs the binding axis while holding safety fixed. We now perform the complementary test: the primary ESMFold2+PRODIGY binding score is held fixed while the safety utility is reconstructed from independent predictors. We rescore toxicity and hemolysis using PeptiVerse's released basic-task peptide-SMILES ChemBERTa CNN and sequence-based XGBoost models, respectively (Zhang et al., 2026b), and obtain a second independent hemolysis score from HemoPI2 Hybrid2 (ESM2-t6+MERCI, model 4) (Rathore et al., 2025). All external scores are complete for the 6,568 candidates used in the primary binding–safety analysis, so every safety definition is evaluated on exactly the same method–target–candidate panel. As in Section C, binding and safety are ranked separately within each method–target stratum before pooling.

Because the safety predictors use different released decision thresholds and score calibrations, their raw outputs are not directly comparable. We therefore define a threshold-normalized signed margin

$$\widetilde{m}(s;\tau) = \begin{cases} (\tau - s)/\tau, & s \le \tau, \\ (\tau - s)/(1-\tau), & s > \tau, \end{cases} \tag{39}$$

which maps the decision boundary to zero and the safe and unsafe score endpoints to $+1$ and -1, respectively; larger values therefore indicate more favorable predicted safety. We use $\tau = 0.5$ for each internal head, 0.49 and 0.2801 for PeptiVerse toxicity and hemolysis, and 0.55 for HemoPI2 Hybrid2.

For each safety definition, the utility is the minimum of its included margins, preserving the benchmark principle that a single adverse liability is sufficient to make a candidate unsafe. ***PeptiVerse substitution*** replaces the internal toxicity and hemolysis margins with their PeptiVerse counterparts, whereas ***HemoPI2 substitution*** replaces only internal hemolysis. ***External consensus*** combines PeptiVerse toxicity and both external hemolysis margins with the unchanged internal allergenicity and neurotoxicity margins. Finally, ***all-oracle consensus*** takes the minimum over all four internal and all three external margins.

Table 6: Binding–safety association under five safety-oracle definitions with the primary binding score held fixed. Safety is represented by the threshold-normalized minimum margin in Equation 39. The stratified estimate ranks candidates within method–target strata before pooling; intervals are 2,000-sample target-bootstrap 95% confidence intervals. The Internal row uses exactly the C.1 candidate panel and recovers its stratified statistic.

Safety definition	n	Strat. ρ	Target-macro ρ [95% CI]	Negative targets
Internal	6,568	-0.224	-0.213 [-0.261, -0.162]	73/82
PeptiVerse substitution	6,568	-0.094	-0.091 [-0.141, -0.043]	56/82
HemoPI2 substitution	6,568	-0.196	-0.173 [-0.223, -0.125]	67/82
External consensus	6,568	-0.096	-0.093 [-0.139, -0.045]	55/82
All-oracle consensus	6,568	-0.116	-0.111 [-0.162, -0.061]	64/82

The central qualitative conclusion is preserved across all five safety definitions: stronger predicted binding is associated with less favorable predicted safety, and every target-macro confidence interval excludes zero. HemoPI2 substitution remains closest to the internal estimate, whereas PeptiVerse substantially attenuates the association without reversing its direction. This result is not induced by the normalization in Equation 39: using the corresponding unnormalized signed margins yields stratified correlations of -0.224, -0.085, -0.202, -0.089, and -0.111 in the same row order, again with all target-macro confidence intervals excluding zero. Thus, the estimated magnitude is safety-oracle dependent, but the negative direction is robust to the independent oracle substitutions considered here.

Table 7 resolves the same analysis by method for the set of methods shown in the binding-oracle comparison. Pepti-drift preserves a negative within-target association under all five safety definitions

($\rho = -0.132$ to -0.086). High-PepBinder and the uniform-random control are likewise consistently negative, whereas estimates close to zero for PepMLM, PepTune, and MOG-DFM can change sign under PeptiVerse-based definitions. These small method-level sign changes do not alter the aggregate target-level conclusion and should not be interpreted as individual significance tests.

Table 7: Method-level within-target binding–safety association as the safety oracle is varied. Entries are Spearman ρ point estimates using the threshold-normalized safety margin over $n = 545$–929 complete candidates per method; no method-level confidence intervals are shown. The displayed methods match Table 5 to facilitate comparison of sensitivity to the binding and safety sides of the evaluation.

Method	Internal	PV sub.	H2 sub.	Ext. cons.	All cons.
Pepti-drift (ours)	-0.132	-0.086	-0.113	-0.093	-0.091
PepMLM	-0.015	0.011	-0.044	-0.017	-0.004
PepTune	-0.091	-0.005	-0.076	-0.014	-0.011
MOG-DFM	-0.110	0.031	-0.094	0.058	0.030
High-PepBinder	-0.400	-0.212	-0.326	-0.204	-0.244
Random uniform-length	-0.444	-0.248	-0.416	-0.256	-0.332

This directional robustness is nontrivial because the safety predictors are far from interchangeable at the candidate level. Within method–target strata, internal and PeptiVerse toxicity scores correlate at -0.158, internal and PeptiVerse hemolysis scores at -0.036, and internal and HemoPI2 hemolysis scores at 0.060. The repeated negative binding–safety association therefore cannot be explained by the external predictors simply reproducing the internal scores. Instead, substantially different safety models disagree about which individual candidates are safest while preserving the same aggregate relationship: stronger predicted binders tend to occupy less favorable regions of predicted safety.

Together with the binding-oracle analysis above, these results show that the direction of the observed binding–safety association survives perturbations on both sides of the computational evaluation pipeline. Its magnitude and the ranking of individual candidates remain oracle-sensitive, so we do not treat any evaluator as biological ground truth. These oracle substitutions show that the unadjusted negative association is not specific to a single evaluator. However, Figure 3 shows that the association is nearly eliminated after exact-length stratification under the primary pipeline. We therefore do not interpret the cross-oracle result as evidence for an intrinsic length-independent binding–safety trade-off. We next show that the same is true of its length mediation.

C.4 Length adjustment of the oracle-robustness analyses

We next test whether the length dependence in Figure 3 is robust to evaluator choice. For each binding- and safety-oracle configuration, we repeat the exact-length analysis by ranking binding and safety within method–target strata and additionally stratifying by exact peptide length, requiring at least three candidates per stratum. Confidence intervals use 2,000 target-cluster bootstrap replicates.

Exact-length stratification eliminates the association in all eight configurations: adjusted correlations range from -0.014 to $+0.032$, and every 95% confidence interval contains zero. The same conclusion holds with unnormalized safety margins, a minimum stratum size of two, and each safety liability analyzed separately (maximum adjusted $|\rho| = 0.060$).

The magnitude of the unadjusted association also tracks each oracle's dependence on peptide length. When safety is fixed, the three binding-oracle estimates follow $\rho(b, L)$; when binding is fixed, the five safety-oracle estimates follow $\rho(u, L)$. Thus, differences in the apparent binding–safety association across evaluators are largely consistent with differences in their length dependence.

Together, these results show that both the unadjusted negative association and its removal by exact-length control are robust to evaluator choice. We therefore interpret the apparent binding–safety trade-off as primarily length-mediated rather than as evidence for a length-independent antagonism between the two objectives.

Table 8: **Length adjustment across binding and safety oracles.** b denotes binding score, u safety utility, and L peptide length. Safety is the minimum signed margin over the four liabilities defining the SAFE–ACTIVE endpoint.

Evaluator	n	Unadjusted ρ	$\rho(b, L)$	$\rho(u, L)$	Exact-length ρ [95% CI]
(a) Binding oracle varied; internal safety fixed					
ESMFold2+PRODIGY	23,993	-0.235	$+0.469$	-0.395	-0.014 [-0.033, $+0.005$]
OpenDDE+EvoEF2	15,596	-0.098	$+0.228$	-0.343	-0.002 [-0.024, $+0.019$]
OpenDDE+HADDOCK3	15,596	-0.137	$+0.305$	-0.343	-0.005 [-0.023, $+0.014$]
(b) Safety oracle varied; ESMFold2+PRODIGY binding fixed					
Internal	6,568	-0.224	$+0.472$	-0.401	$+0.010$ [-0.030, $+0.048$]
PeptiVerse substitution	6,568	-0.094	$+0.472$	-0.166	$+0.014$ [-0.031, $+0.058$]
HemoPI2 substitution	6,568	-0.196	$+0.472$	-0.363	-0.008 [-0.045, $+0.029$]
External consensus	6,568	-0.096	$+0.472$	-0.174	$+0.022$ [-0.025, $+0.067$]
All-oracle consensus	6,568	-0.116	$+0.472$	-0.207	$+0.032$ [-0.011, $+0.073$]

Table 9: **Full fixed-budget joint evaluation at $K = 30$.** SA denotes the percentage of all allocated candidate slots satisfying the complete SAFE–ACTIVE endpoint. Enrichment is the percentage-point improvement over the target- and exact-length-matched random control; brackets show target-bootstrap 95% confidence intervals. SA target coverage is the percentage of held-out targets with at least one SAFE–ACTIVE hit.

Method	SA ↑	Enrich. ↑ [95% CI]	SA target cov. ↑	peptides/sec ↑
Pepti-drift (ours)	<u>18.37</u>	$+2.65$ [0.08, 5.42]	60.23	**378.3**
PepTune	**18.45**	-0.49 [-2.46, 1.52]	67.05	0.81
MOG-DFM	10.98	-0.53 [-2.50, 1.40]	<u>71.59</u>	0.01
High-PepBinder	10.11	-2.05 [-3.22, -0.83]	65.91	0.73
PepMLM	8.83	$+\mathbf{5.11}$ [$\mathbf{1.97}$, $\mathbf{8.67}$]	51.14	82.23
AR Transformer	14.55	$+0.83$ [-0.68, 2.35]	**72.73**	<u>340.2</u>

D ADDITIONAL FIXED-BUDGET DIAGNOSTICS

D.1 FULL FIXED-BUDGET JOINT RESULTS

Table 9 reports the numerical results underlying the main-text quality–cost comparison. All methods are evaluated under the same fixed budget of $K = 30$ candidates per target over 88 held-out targets. Raw SAFE–ACTIVE yield measures absolute joint success, whereas Enrichment@30 reports the percentage-point improvement over the target- and exact-length-matched random control. Target coverage is the fraction of targets with at least one SAFE–ACTIVE candidate. Generation throughput is measured under the protocol described in Appendix F.4.

D.2 MARGINAL SAFETY AND BINDING RESULTS

The main comparison focuses on whether the same generated candidate satisfies both binding and safety constraints. To clarify how this joint performance arises, Tables 10 and 11 report the corresponding marginal safety and binding results. These quantities are descriptive rather than controlled leaderboards: the primary comparison remains the fixed-budget SAFE–ACTIVE endpoint together with exact-length-matched Enrichment@30.

D.3 HIGH-DEPTH TARGET ANALYSIS

The main-text analysis summarizes the distribution of SAFE–ACTIVE discovery depth among reached targets. We further examine performance at increasingly stringent depth thresholds by defining

$$N_{\geq n} = \sum_{t \in \mathcal{T}} \mathbf{1}[D_t \geq n],$$

Table 10: Predicted-safety results at $K = 30$. All entries are percentages. Safe-only is the fraction of submitted candidates satisfying sequence-integrity requirements and all four safety gates without applying the binding criterion. Target safe is the fraction of targets with at least one such candidate. Tox, Hemo, All., and Neuro denote the fractions passing the toxicity, hemolysis, allergenicity, and neurotoxicity gates, respectively. Bold marks the best value and underline marks the second-best value among the six noncollapsed learned or guided methods.

Method	Safe-only	Target safe	Tox	Hemo	All.	Neuro
Pepti-drift (ours)	<u>61.17</u>	<u>98.86</u>	<u>90.61</u>	<u>94.55</u>	80.64	<u>82.23</u>
PepMLM	34.09	88.64	72.20	88.41	69.39	62.77
PepTune	**72.73**	**100.00**	**96.86**	**96.14**	**88.75**	**90.83**
AR Transformer	51.93	**100.00**	84.13	91.48	78.18	75.49
MOG-DFM	57.90	**100.00**	88.76	90.02	<u>88.42</u>	82.04
High-PepBinder	42.88	**100.00**	78.41	88.18	78.52	64.62
Random length-matched	45.23	100.00	82.31	77.08	84.24	68.22
Random uniform-length	35.68	100.00	71.33	70.11	84.85	57.39
Train prior	0.00	0.00	85.83	92.65	78.41	78.11

Table 11: Predicted-binding results at $K = 30$. Bind cov. is the percentage of submitted candidates receiving a binding score after structure-confidence gating. Bind-active additionally requires $s_{\mathrm{bind}} \geq 7$; both use all submitted candidates as the denominator (2,640 per method; 2,634 for MOG-DFM). Target pass thresholds each target's mean ipTM, interface pLDDT, and $-\Delta G$ at 0.5, 70, and 7, respectively. These means use all complete measurement triples before confidence gating; all 88 targets are evaluable. This diagnostic is distinct from any-hit target coverage. Mean scored averages available confidence-gated binding scores; ipTM and pLDDT average the corresponding target means. Bold and underline indicate the best and second-best values among the six noncollapsed learned or guided methods.

Method	Bind cov.	Bind-active	Target pass	Mean scored	ipTM	pLDDT
Pepti-drift (ours)	**35.45**	<u>25.80</u>	**26.14**	8.20	**0.442**	<u>84.44</u>
PepMLM	31.10	**27.16**	<u>25.00</u>	**9.51**	<u>0.425</u>	82.79
PepTune	<u>32.54</u>	24.32	<u>25.00</u>	8.39	<u>0.425</u>	**85.72**
AR Transformer	29.13	23.37	25.00	8.58	0.416	83.86
MOG-DFM	23.46	18.75	15.91	8.68	0.372	83.35
High-PepBinder	25.15	19.36	17.05	<u>8.82</u>	0.386	83.27
Random length-matched	26.33	21.82	18.18	9.05	0.398	83.78
Random uniform-length	20.64	16.74	9.09	9.17	0.370	82.51
Train prior	25.80	20.15	19.32	8.79	0.391	83.64

the number of held-out targets for which at least n of the $K = 30$ submitted candidates satisfy the complete SAFE–ACTIVE endpoint.

Figure 8 shows $N_{\geq n}$ for depth thresholds $n = 15, \ldots, 30$. PepTune initially reaches one more target at $n \geq 15$ (11 versus 10), and the two methods are tied at $n \geq 18$ with nine targets each. At more stringent thresholds, however, the separation increasingly favors Pepti-drift. At $n \geq 21$, Pepti-drift retains seven targets, compared with four for PepTune and two for PepMLM. At $n \geq 24$, Pepti-drift retains five targets, whereas PepTune retains none and PepMLM retains one.

At the highest depth thresholds, Pepti-drift is the only method with any target at $n \geq 26$, retains four targets at $n \geq 27$, and reaches $30/30$ SAFE–ACTIVE candidates on one target. Thus, the larger conditional mean observed in the main text reflects consistently deeper feasible candidate pools across multiple targets rather than being driven by a single exceptional target.

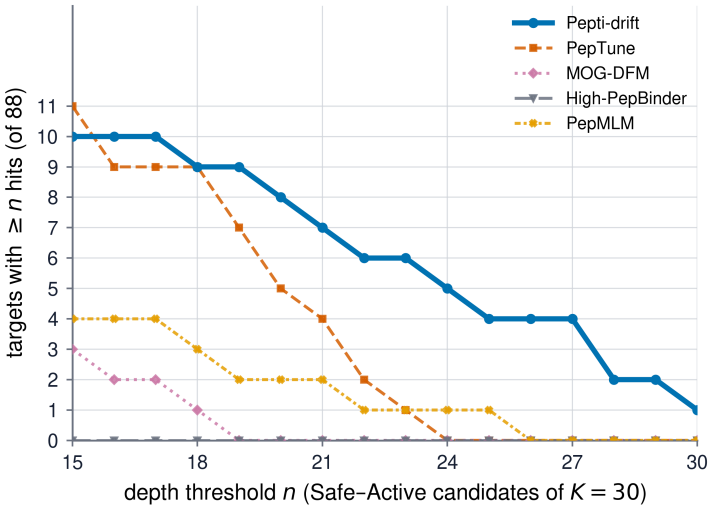


Figure 8: **Target-level discovery at high SAFE–ACTIVE candidate counts.** For each threshold n, the plot reports the number of the 88 held-out targets for which at least n of $K = 30$ submitted candidates satisfy the complete SAFE–ACTIVE endpoint. Pepti-drift separates from the other methods at high candidate-count thresholds and is the only method with any target containing at least 26 SAFE–ACTIVE candidates.

D.4 MECHANICALLY SELECTED TARGET-LEVEL EXAMPLES

To complement the aggregate discovery-depth analysis, we provide three target-level examples selected by prespecified rules from the frozen SAFE–ACTIVE counts. These examples are intended to illustrate distinct target-level behaviors rather than to define the aggregate comparison.

For *rescue* (rule R), we select a target reached by Pepti-drift but by none of the other four published peptide generators considered here. This rule identifies 2b1n_A, on which Pepti-drift produces 2/30 SAFE–ACTIVE candidates while all four comparators produce none.

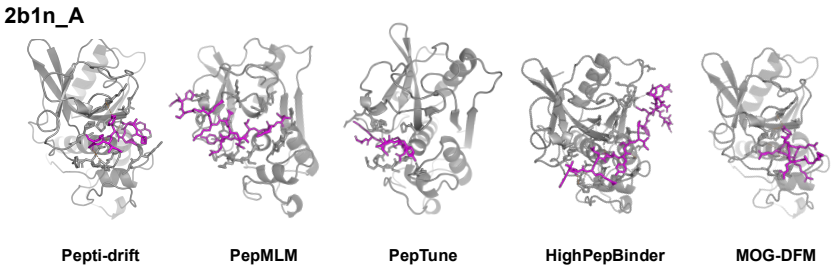
For *depth* (rule D), we restrict attention to the 26 targets reached by all five methods and rank them by Pepti-drift's depth advantage, defined as its SAFE–ACTIVE count minus that of the strongest comparator on the same target. The two largest gaps occur on 3cvl_A and 2qa9_E, where the advantages are $+12$ and $+11$ candidates, respectively. On 3cvl_A, all 30 Pepti-drift submissions satisfy the complete joint endpoint, compared with 18/30 for the strongest comparator; on 2qa9_E, the corresponding counts are 23/30 and 12/30.

Figure 9 visualizes these mechanically selected examples. Within each target–method pair, the displayed peptide is selected independently of the SAFE–ACTIVE count by maximum finite ESMFold2 ipTM, with submitted rank used only to break ties.

D.5 PREDICTED-STRUCTURE INTERPRETATION OF HIGH-DEPTH EXAMPLES

We further examined the predicted structures of the two depth targets mechanically selected in Appendix D.4, 3cvl_A and 2qa9_E. Importantly, these targets were selected from the frozen SAFE–ACTIVE counts before inspection of the structural descriptors. Within each target–method pair, we use the same maximum-finite-ipTM representative defined in Appendix D.4. This analysis is intended to interpret the benchmark predictions rather than to provide independent validation of binding or biological safety.

Structural descriptors. A target–peptide residue contact was defined when any inter-chain heavy-atom pair was within 4.5 Å, with each residue pair counted once. The conclusions were additionally checked at 4.0, 5.0, and 5.5 Å. Solvent-accessible surface area (SASA) was calculated with FreeSASA using ProtOr radii and a 1.4 Å probe. Peptide burial was defined as the SASA of the isolated peptide in its bound conformation minus its SASA in the complex. We separately quantified buried and exposed



Method	Peptide	$\Delta G \downarrow$	ipTM ↑	Toxicity ↓	Hemolysis ↓	Safe–Active hits/submitted
Target 2b1n_A (rule R)						
Pepti-drift (ours)	SAYNYIAV	−10.5	0.518	0.001	0.019	**2/30**
MOG-DFM	REEERRY	−8.1	0.448	**0.000**	**0.002**	0/29
PepTune	VYNGCRS	−8.9	0.650	0.011	0.260	0/30
High-PepBinder	GTPRLEYSIPHLNDDGETNN	**−12.5**	0.682	0.923	0.013	0/30
PepMLM	LKTTTYEENKQGSGH	−11.4	**0.685**	0.810	0.012	0/30
Target 3cvl_A (rule D, +12)						
Pepti-drift (ours)	RFNPF	−9.2	0.938	**0.000**	**0.001**	**30/30**
MOG-DFM	YWKERIT	−9.4	0.874	0.037	0.052	10/29
PepTune	VYFSKLD	−9.6	0.922	0.007	0.020	18/30
High-PepBinder	LCSAID	**−10.3**	**0.941**	0.000	0.277	9/30
PepMLM	QFSEELLLRLALKKL	−9.5	0.778	0.264	0.065	4/30
Target 2qa9_E (rule D, +11)						
Pepti-drift (ours)	ARPI	−10.7	**0.958**	**0.000**	**0.119**	**23/30**
MOG-DFM	LLRYPLWRQTAWEGW	**−15.3**	0.882	0.071	0.902	10/30
PepTune	ISKLRDD	−11.3	0.933	0.028	0.224	12/30
High-PepBinder	IHSLW	−11.6	0.939	**0.000**	0.125	9/30
PepMLM	DCVTYYIPPPAGLLK	−13.6	0.804	0.357	0.527	1/30

Figure 9: **Mechanically selected examples of target-level rescue and discovery depth.** Targets are selected from frozen per-target SAFE–ACTIVE counts using the prespecified rules described in Appendix D.4. Rule R identifies 2b1n_A, which is reached by Pepti-drift (2/30 hits) but by none of the other four methods shown. Rule D selects the two largest Pepti-drift depth advantages among the 26 targets reached by all five methods: 3cvl_A (30/30 versus 18/30 for the strongest comparator) and 2qa9_E (23/30 versus 12/30). The structural examples above the table show target 2b1n_A; within each target–method pair, the displayed peptide is selected mechanically by maximum finite ESMFold2 ipTM, with submitted rank breaking ties. The final column reports SAFE–ACTIVE hits over the complete submitted set. Bold indicates the best value within each target.

hydrophobic side-chain surface for AVILMFWYPC residues. Because contact and surface-area quantities depend strongly on peptide length, extensive quantities are reported per peptide residue.

3cvl_A: productive hydrophobic burial. On 3cvl_A, Pepti-drift's fixed representative RFNPF (ipTM = 0.938, predicted $\Delta G = -9.2$ kcal/mol, hemolysis score = 0.001) has the highest contact density among the five fixed representatives (6.40 target-residue contacts per peptide residue) and the largest peptide buried area (137.8 Å^2 per residue). It buries 79.0% of its hydrophobic side-chain surface while leaving 17.7 Å^2 per residue exposed.

This pattern is not restricted to the single displayed candidate. Across all submitted Pepti-drift candidates on 3cvl_A, the median contact density is 5.00 per residue, the median buried peptide area is 106.2 Å^2 per residue, and the median buried hydrophobic-side-chain SASA is 76.0 Å^2 per residue. The latter is approximately 2.2-fold higher than the largest competitor median. Relative to Pepti-drift's same-length random controls, buried hydrophobic-side-chain SASA increases by 24.9 Å^2 per residue on average, with 28 of 30 paired differences positive. Thus, the high SAFE–ACTIVE depth

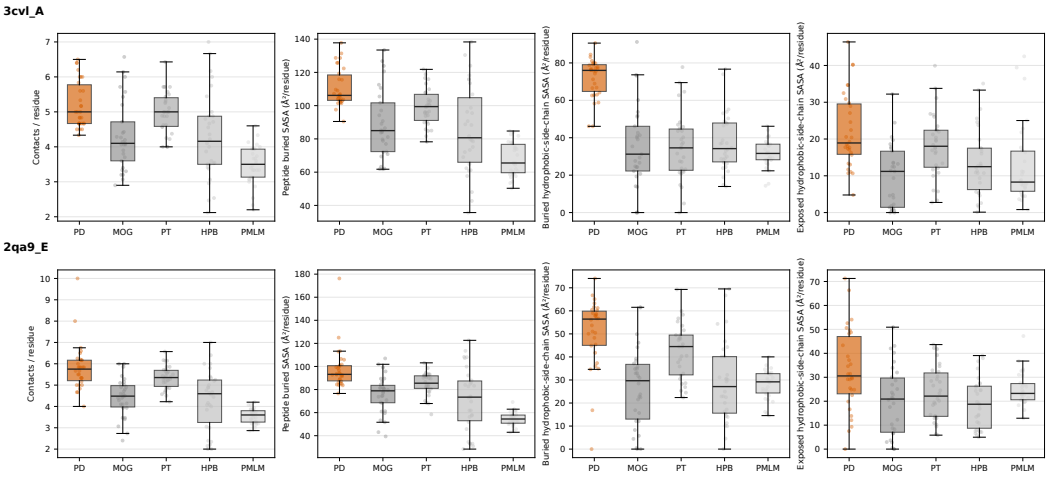


Figure 10: **Targets with large Pepti-drift candidate pools exhibit stronger predicted interface utilization.** Distributions across all submitted candidates are shown for 3cvl_A and 2qa9_E, the two targets with the largest Pepti-drift SAFE–ACTIVE candidate-count advantage among targets reached by all five methods. Panels show contacts per residue, buried peptide SASA, buried hydrophobic-side-chain SASA, and exposed hydrophobic-side-chain SASA, with surface areas normalized per peptide residue. Boxes show the interquartile range and median; points denote individual candidates. PD, Pepti-drift; MOG, MOG-DFM; PT, PepTune; HPB, High-PepBinder; PMLM, PepMLM.

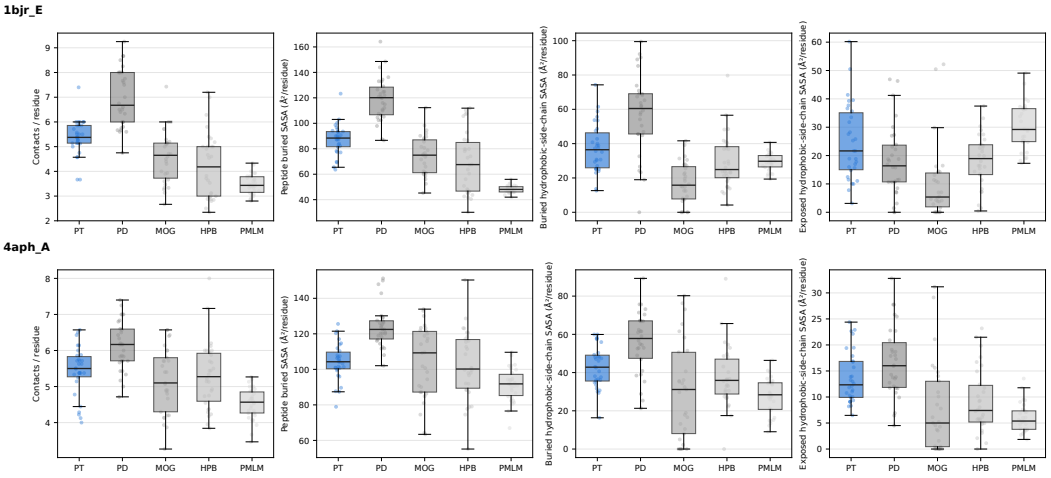


Figure 11: **Pepti-drift exhibits stronger predicted interface utilization on PepTune's top SAFE-ACTIVE targets.** Distributions across all submitted candidates are shown for 1bjr_E and 4aph_A, the two targets with the largest SAFE–ACTIVE candidate counts for PepTune. Panels show contacts per residue, buried peptide SASA, buried hydrophobic-side-chain SASA, and exposed hydrophobic-side-chain SASA, with surface areas normalized per peptide residue. Boxes show the interquartile range and median; points denote individual candidates. PT, PepTune; PD, Pepti-drift; MOG, MOG-DFM; HPB, High-PepBinder; PMLM, PepMLM.

on this target is accompanied by substantial productive interfacial burial rather than being explained by peptide length alone.

2qa9_E: spatial allocation rather than lower sequence hydrophobicity. On 2qa9_E, Pepti-drift's ARPI provides an illustrative contrast with the MOG-DFM representative LLRYPLWRQTAWEGW. MOG-DFM has the stronger predicted binding energy (-15.3 versus -10.7 kcal/mol) but a substantially less favorable hemolysis score (0.902 versus 0.119). In the predicted complexes, ARPI

Table 12: **Structural descriptors of mechanically selected high-depth examples.** Shown are the two targets selected by the prespecified discovery-depth rule, 3cvl_A and 2qa9_E. For each target–method pair, the representative peptide is the fixed maximum-ipTM candidate. Contact density denotes the number of target-residue contacts per peptide residue; buried SASA and exposed hydrophobic side-chain SASA are normalized by peptide length; Hydro. SC burial denotes the fraction of hydrophobic side-chain surface buried upon complex formation. The final column reports the number of SAFE–ACTIVE candidates among all submitted candidates for that target. These descriptors provide structural interpretation of the predicted complexes rather than independent binding validation.

	3cvl_A					2qa9_E				
Method	Contact density ↑	Buried SASA (Å^2/res) ↑	Hydro. SC burial ↑	Exposed hydro. SASA (Å^2/res) ↓	Safe–Active hits/submitted ↑	Contact density ↑	Buried SASA (Å^2/res) ↑	Hydro. SC burial ↑	Exposed hydro. SASA (Å^2/res) ↓	Safe–Active hits/submitted ↑
Pepti-drift (ours)	**6.40**	**137.8**	**79%**	**17.7**	**30/30**	**6.25**	**112.8**	**82%**	13.8	**23/30**
MOG-DFM	6.00	123.9	62%	24.4	10/29	3.93	59.5	49%	36.5	10/30
PepTune	5.71	115.9	56%	33.8	18/30	5.71	93.1	76%	**9.4**	12/30
High-PepBinder	6.17	115.6	66%	25.6	9/30	6.00	108.3	60%	36.7	9/30
PepMLM	2.53	56.9	54%	25.0	4/30	4.07	62.2	47%	36.8	1/30

shows higher contact density (6.25 versus 3.93 contacts per residue), greater buried peptide area (112.8 versus 59.5 Å^2 per residue), and a larger hydrophobic-side-chain burial fraction (81.6% versus 48.8%), while leaving less hydrophobic-side-chain area exposed (13.8 versus 36.5 Å^2 per residue).

Notably, ARPI is not less hydrophobic at the sequence level: its GRAVY score is 0.05 compared with -0.63 for the MOG-DFM representative, and its hydrophobic-residue fraction is also higher (0.75 versus 0.60). This example therefore supports an interpretation based on *spatial allocation of hydrophobicity*: hydrophobic residues can be retained but productively buried at the predicted target interface rather than simply depleted from the sequence.

Scope of interpretation. The reduced exposed hydrophobic surface observed for the fixed representatives does not generalize uniformly across all submitted candidates. In particular, Pepti-drift does not have the lowest median exposed hydrophobic-side-chain SASA on both targets. We therefore interpret the consistent signal as strong *interface utilization* and *productive hydrophobic burial*, rather than a general tendency to minimize exposed hydrophobicity. Moreover, these descriptors are calculated from the same ESMFold2 complexes used by the benchmark binding pipeline and should therefore be viewed as qualitative structural interpretation, not independent evidence of binding or a causal mechanism for predicted safety.

D.6 INTEGRITY AND OMITTED CONTROLS

Pepti-drift generates 100% canonical, in-range sequences with 100% within-target uniqueness, 99.51% exact novelty relative to its generator training peptides, and no exact cross-target reuse.

The integrity gates are important for several architecture controls. Conditional discrete diffusion, the conditional VAE, and the non-autoregressive Transformer have within-target uniqueness rates of 3.33%, 1.33%, and 0.15%, respectively, and all three achieve zero SAFE–ACTIVE yield and zero target coverage. The non-autoregressive Transformer emits only four distinct sequences among 2,640 outputs despite nominally favorable values on some marginal properties. The training-peptide prior likewise achieves zero SAFE–ACTIVE yield because all candidates fail exact novelty.

The exact-length-matched random control reaches 10.27% raw SAFE–ACTIVE yield, 69.32% target coverage, and 26.33% confidence-gated binding coverage. Its relatively strong raw yield further illustrates why raw SAFE–ACTIVE rate alone is not sufficient for controlled comparison.

D.7 THRESHOLD SENSITIVITY

We re-gate the frozen $K = 30$ candidate scores for all 12 methods and their method-specific, same-length random controls. No predictor or structure model is rerun, and candidates with missing scores continue to fail the gate. On the joint 5×5 grid, the median Spearman correlation with the default method ranking is 0.928 (minimum 0.840). PepMLM ranks first in 21 of 25 settings, whereas Pepti-drift retains positive enrichment in 24 settings and remains in the top three throughout.

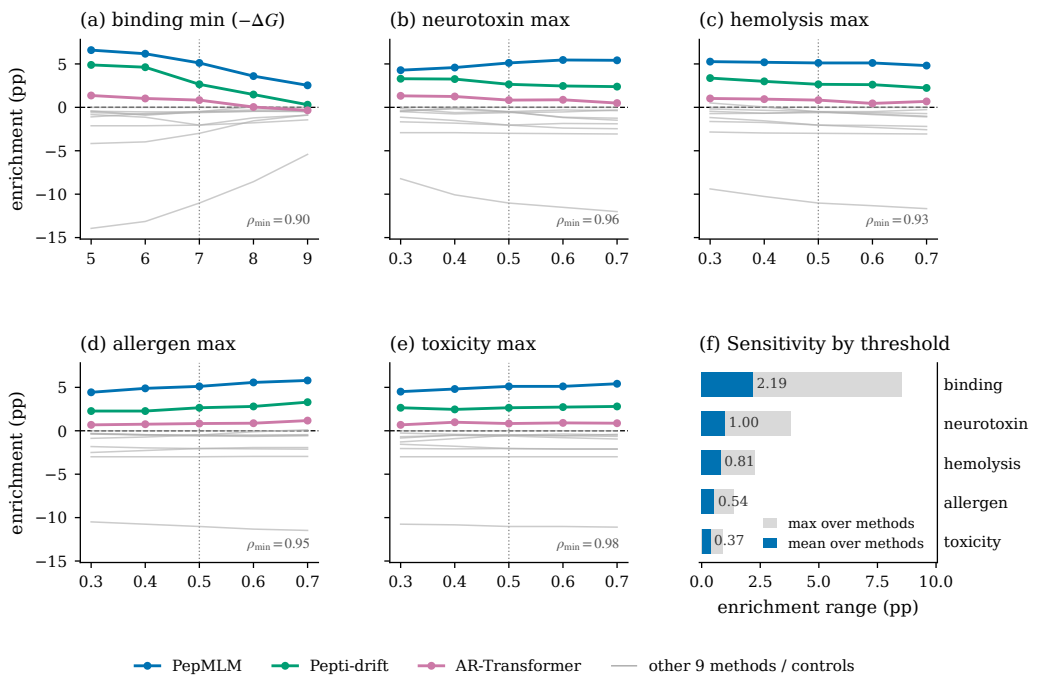


Figure 12: **One-at-a-time sensitivity of Safe-Active Enrichment@K.** Frozen scores are re-gated on 88 test targets while varying one threshold and holding the other four at the benchmark default. **(a)** Binding-score minimum. **(b–e)** Upper thresholds for neurotoxicity, hemolysis, allergenicity, and toxicity. Coloured curves show the three methods with positive enrichment at the default gate; grey curves show the other nine methods and controls. Dotted lines mark the default threshold, and $\rho_{\min}$ is the lowest Spearman correlation with the default 12-method ranking in each sweep. **(f)** The enrichment range induced by each threshold, averaged over methods (blue) and for the most affected method (grey).

To identify which gate drives this variation, we additionally sweep one threshold at a time while fixing the other four at their defaults. The binding minimum takes values $\{5, 6, 7, 8, 9\}$; each upper safety threshold takes values $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. Unlike the joint sweep, this analysis separates the effects of toxicity, hemolysis, allergenicity, and neurotoxicity.

Figure 12 shows that the binding cutoff is the dominant source of threshold sensitivity. Increasing the binding minimum from 5 to 9 reduces enrichment for each of the three leading methods; for Pepti-drift it falls from $+4.89$ to $+0.34$ percentage points. Nevertheless, the worst rank correlation within this sweep remains high ($\rho_{\min} = 0.900$). Averaged over methods, the binding sweep induces a 2.19 percentage-point enrichment range, compared with 1.00 for neurotoxicity, 0.81 for hemolysis, 0.54 for allergenicity, and 0.37 for toxicity. Correspondingly, the individual safety sweeps retain rank correlations of at least 0.928, with toxicity being nearly inert ($\rho_{\min} = 0.977$).

PepMLM remains the highest-enrichment method at every value in each one-at-a-time sweep, and Pepti-drift remains positively enriched throughout. The safety curves need not be monotone because enrichment is a paired difference: relaxing a safety gate can admit additional candidates from both the model and its same-length random control. Together with the joint-grid results, these sweeps show that the qualitative comparison is not determined by one default cutoff. They remain conditional on the frozen evaluators, however, and cannot address evaluator misspecification or experimental validity.

E PEPTI-DRIFT ARCHITECTURE AND TRAINING CONFIGURATION

Table 13 summarizes the architecture, training, and optimization configuration of the main Pepti-drift checkpoint used in the benchmark. The fixed peptide reference map follows the construction described in Appendix A.2.

F REPRODUCIBILITY DETAILS

F.1 COMPARATOR SUBMISSIONS

The eight comparators to Pepti-drift are PepMLM, PepTune, High-PepBinder, MOG-DFM, conditional autoregressive and non-autoregressive Transformers, a conditional VAE, and conditional discrete diffusion. The four bundled neural baselines use mean-pooled ESM-2 3B target embeddings and the same strict binding training export as Pepti-drift. We additionally evaluate a training-peptide prior, random peptides sampled from the training length distribution, and random peptides with length uniformly sampled from 4–25.

PepTune is restricted before submission to linear peptides containing the 20 standard L-amino acids. MOG-DFM generates ten candidates at each of length quantiles 7, 10, and 15; six duplicated outputs leave 2,634 observed slots, and the missing slots remain failures in fixed-budget metrics. High-PepBinder uses its released checkpoint and preserves generation order. Unless a method underfills its budget, each run contains 2,640 candidates.

The main leaderboard contains one declared generation run per method. Accordingly, confidence intervals in the fixed-budget comparison quantify variation across held-out targets rather than variation from repeated model training.

F.2 BINDING-EVALUATOR IMPLEMENTATION

For complex-structure prediction in BindSafe-PepBench, we use ESMFold2. Specifically, we use the `Synthyra/ESMFold2-Experimental-Fast-Cutoff2025` packaging of the corresponding Biohub checkpoint. Interface binding energies are then computed from the predicted complexes with PRODIGY (`prodigy-prot` ≥ 2.4) (Xue et al., 2016). All methods are evaluated with the same checkpoint and inference configuration. Because the ESMFold2 repository is an actively updated remote-code model tracking `main`, the checkpoint is resolved once and reused for every method rather than refetched per run.

F.3 PRETRAINED WEIGHTS AND EXTERNAL CHECKPOINTS

Every pretrained checkpoint used anywhere in the benchmark is fixed for all methods and listed below. Weights are resolved from the Hugging Face Hub unless stated otherwise.

- **Complex-structure prediction (primary binding oracle).** `Synthyra/ESMFold2-Experimental-Fast-Cutoff2025`, as described above.
- **Target encoder for Pepti-drift and the bundled baselines.** ESM-2 3B, `facebook/esm2_t36_3B_UR50D`, mean-pooled over the target sequence. The same embeddings are used by Pepti-drift, the conditional autoregressive and non-autoregressive Transformers, the conditional VAE, and conditional discrete diffusion; these five models carry no other pretrained weights and are trained in this work from the strict binding training export.
- **Safety and property evaluators.** ESM-2 650M, `facebook/esm2_t33_650M_UR50D`, used as a frozen mean-pooling encoder under a lightweight MLP head trained in this work for each of the seven properties (toxicity, hemolysis, allergenicity, neurotoxicity, nonfouling, cell penetration, and hemolytic HC_{50}). The four properties defining the SAFE–ACTIVE gate reach held-out ROC-AUC of 0.756 (toxicity), 0.825 (hemolysis), 0.874 (allergenicity), and 0.700 (neurotoxicity).
- **PepMLM.** `TianlaiChen/PepMLM-650M`, the authors' released checkpoint.
- **MOG-DFM.** The gated `ChatterjeeLab/MOG-DFM` release, comprising the peptide flow-matching model and the authors' guidance classifiers for binding affinity, hemolysis,

nonfouling, and half-life. Its internal sequence encoder is ESM-2 650M, the same checkpoint as above.

- **PepTune.** The authors' released masked discrete diffusion checkpoint together with their SMILES SPE tokenizer, PeptideCLM embedder, and their own property classifiers. PepTune runs in its own container and is the only comparator whose scores originate from a SMILES-level model.
- **High-PepBinder.** The released latent-diffusion checkpoint `pepldm_step=882-val_loss=0.51.ckpt` (5.69 GB, Zenodo record `18282817`, MD5 `e98b1fc12a0f0c73f59775196e9428f3`), used with its released batch size of 64.
- **Independent binding oracles.** The OpenDDE v1 checkpoint `opendde.pt` generates the poses shared by the OpenDDE+EvoEF2 and OpenDDE+HADDOCK3 tracks of Appendix C.2; EvoEF2 and HADDOCK3 are scoring functions and carry no learned weights.

The training-peptide prior and the two random-peptide controls use no pretrained weights.

F.4 Generation-speed protocol

Generation time is measured separately for each of the 88 held-out targets on one NVIDIA H100 80GB HBM3, from the raw target amino-acid sequence to the decoded peptide strings. The timed region includes target preprocessing and embedding required by the method, sampling or inference-time property guidance, decoding, native validity filtering, and any additional sampling needed to reach the requested budget. Model and tokenizer loading, checkpoint loading, disk I/O, and one target-independent warm-up are excluded.

During leaderboard generation, Pepti-drift and the autoregressive Transformer can read precomputed ESM-2 target embeddings. For the timing comparison, both instead re-embed every target inside the timed region so that their target encoding cost is treated consistently with the other sequence-conditioned methods. Target embedding accounts for approximately 80% of Pepti-drift's measured runtime and 72% of the autoregressive Transformer's runtime.

Pepti-drift, PepMLM, the autoregressive Transformer, and High-PepBinder were measured under the common timing harness. The three fastest methods were averaged over five complete passes over the 88 targets. PepTune and MOG-DFM use the per-target clocks recorded by their original $K = 30$ generation runs under the same model-loading and I/O exclusions. MOG-DFM was not rerun solely for timing because doing so would require approximately 61 additional GPU-hours; its original timer also used wall-clock time without an explicit CUDA synchronization. These implementation differences should be considered when interpreting precise runtime ratios, although they do not affect the orders-of-magnitude separation from Pepti-drift.

The measured costs also reflect constraints required by the benchmark. PepTune continues sampling until 30 parseable linear canonical peptides are obtained. MOG-DFM assembles $K = 30$ from three separate length-specific generation calls. High-PepBinder uses its released batch size of 64 and therefore oversamples the requested budget of 30.

Table 13: **Architecture and training hyperparameters for Pepti-drift.** Values correspond to the main checkpoint used in the benchmark.

Category	Hyperparameter	Value
Representation	ESM-2 encoder	`esm2_t36_3B_UR50D`
	ESM embedding dimension	2560
	ESM pooling	mean pooling
	ESM parameters	frozen
	Target projection	$2560 \to 2560 \to 2560$
	Target projection activation	ReLU
	Peptide reference projection	$2560 \to 1024 \to 1024$
	Peptide reference activation	ReLU
	Peptide reference parameters	fixed after random initialization
	Latent dimension d_z	1024
	Latent normalization	ℓ_2 normalization
Generator	Architecture	$3584 \to 2048 \to 2048 \to 1024$
	Hidden activation	ReLU
	Noise distribution	$\mathcal{N}(0, I_{1024})$
Drift module	Architecture	$3584 \to 2048 \to 2048 \to 1024$
	Hidden activation	ReLU
	Number of drift steps	1
	Drift step size η	1.0
	Output gate	learned scalar, initialized to 1.0
Decoder	Architecture	non-causal Transformer
	Maximum token slots L	50
	Model dimension	1024
	Transformer layers	4
	Attention heads	8
	Feed-forward dimension	4096
	Transformer activation	GELU
	Transformer normalization	pre-norm
	Dropout	0.1
	Positional representation	learned queries
	Vocabulary	20 amino acids + special tokens
Attraction–repulsion	Attraction coefficient α_+	1.0
	Sampled negatives per example	64
	Retained hard negatives K_{neg}	32
	Kernel temperature τ	0.1
	Final repulsion coefficient α_-	1.0
	Repulsion start e_{start}	100
	Repulsion warm-up e_{warmup}	500 epochs
	Warm-up schedule	linear
	Initial repulsion	0
Objective	λ_{seq}	1.0
	λ_{drift}	1024
	λ_{reg}	10.24
Optimization	Optimizer	AdamW
	Learning rate	1×10^{-4}
	Weight decay	1×10^{-4}
	Learning-rate scheduler	none
	Batch size	32
	Maximum epochs	1000
	Early-stopping patience	10 epochs
	Early-stopping tolerance	1×10^{-4}
	Gradient clipping	none
	Random seed	42

Table 14: Generation speed for producing $K = 30$ candidates per held-out target on one NVIDIA H100. Per-candidate cost divides by the requested budget of 30 rather than by the number returned, so under-delivery is not rewarded. Relative cost is normalized to Pepti-drift.

Method	Time/target (s)	Time/candidate (ms)	Candidates/s	Relative
MOG-DFM	2,492.0	83,065.8	0.012	31,426.8×
High-PepBinder	40.85	1,361.5	0.734	515.1×
PepTune	37.08	1,236.0	0.809	467.6×
PepMLM	0.365	12.16	82.23	4.6×
AR Transformer	0.088	2.940	340.2	1.1×
Pepti-drift (ours)	**0.079**	**2.643**	**378.3**	1.0×