APEXMO: EDIT-CONSTRAINED MULTI-OBJECTIVE OPTIMIZATION OF GENERATIVE PEPTIDE PROPOSALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative sequence models produce diverse peptide proposals, but refining a proposal requires balancing competing properties while controlling how far its sequence changes. We present ApexMO, a computational framework that combines length-conditioned discrete flow matching with anchor-relative, edit-constrained Pareto optimization. Generated peptides provide fixed anchors for local refinement of predicted antimicrobial activity, toxicity, and hemolysis. Under matched raw-attempt budgets, the flow-matching generator yielded approximately 1.8 times as many sequences meeting the specified novelty, predicted-activity, and ensemble-dispersion criteria as each of the two tested baseline systems. Across five refinement runs, all-survivor hypervolume of predicted activity, toxicity, and hemolysis increased by 26.8–27.9%. Under equal evaluation budgets, Pareto-based search achieved higher hypervolume than random local search, both for the final population and for all feasible solutions evaluated during a run. Adding crossover did not improve the median relative to mutation-only search. An edit-budget ablation quantified the trade-off between predicted objective improvement and anchor proximity. A screening cascade prioritized 35 DBAASP-predicted AMP candidates from 25 generated anchors, with recorded sequence ancestry. These results are based on model predictions and do not establish experimental antimicrobial activity or safety.

1 INTRODUCTION

Antimicrobial peptides (AMPs) must balance antibacterial activity with mammalian toxicity, hemolysis, and other properties that can conflict with potency (Lazzaro et al., 2020; Oliveira Júnior et al., 2025; Wan et al., 2024b). A generator can explore sequence space, but its samples do not by themselves specify which trade-offs should be retained. Conversely, optimizing a set of predictive scores can improve the nominal objectives while moving away from the starting scaffold or exploiting inaccuracies in the predictors.

We study a two-stage setting: a learned prior supplies peptide proposals, and a local optimizer refines them subject to an edit budget around each original proposal. The optimizer returns multiple activity–safety compromises rather than a single weighted score optimum. Every descendant retains a fixed anchor and a recorded ancestry. This separates three properties that are often conflated: provenance records where a sequence came from; locality limits how far it can change; and predictive validity determines whether score improvements are likely to transfer beyond the optimization models. Our experiments directly assess the first two and report the limitations of the third.

ApexMO combines a length-conditioned discrete flow-matching prior (Gat et al., 2024) with constrained non-dominated sorting genetic algorithm II (NSGA-II) search (Deb et al., 2002). The learned generator is used for global proposal generation; it is not retrained during local search. The search minimizes predicted minimum inhibitory concentration (MIC), toxicity, and hemolysis while enforcing anchor-relative Levenshtein bounds, novelty requirements, predictor applicability domains, and an activity-uncertainty threshold. A subsequent cascade selects non-redundant candidates and records their sequence-level relationship to the generated anchors.

We integrate these established methods into a design framework and evaluate the tested generator systems, local-search controls, and edit-budget settings. We address three questions. First, does the learned prior generate AMP-associated sequences with useful predicted activity and novelty

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.

under a fixed sampling budget? Second, does constrained multi-objective selection improve the final population relative to matched local-search controls? Third, how does the edit budget change objective improvement, the number of selected representatives, and anchor proximity? The results show consistent surrogate-space improvements and quantify the cost of tighter local constraints. Adding crossover did not improve the median final all-survivor hypervolume under the tested budget. The final candidates remain computational proposals for experimental evaluation.

2 Related work

Generative peptide design. HydrAMP, latent diffusion models, and discrete diffusion models demonstrate several routes to generating peptide candidates (Szymczak et al., 2023; Torres et al., 2025; Markakis et al., 2026). Discrete flow matching provides a general framework for learning distributions over discrete sequences (Gat et al., 2024). We use its masked-source construction as an AMP prior and compare a fixed checkpoint of this prior with AMP-Diffusion and ARCADIAMP under matched raw-attempt budgets. Because training corpora and sampling configurations differ, this comparison evaluates the tested systems rather than isolating an architectural advantage of flow matching.

Sequence optimization and local constraints. Virtual evolution and template-based optimization improve proposed or established peptide scaffolds (Wang et al., 2025; Torres et al., 2026). ApexGO already imposes template-similarity constraints and updates its latent generator during Bayesian optimization (Torres et al., 2026). LaMBO jointly learns a denoising model and multi-task surrogate for multi-objective optimization (Stanton et al., 2022); MosPro uses pretrained property predictors for Pareto-guided local sampling (Luo et al., 2025). PepTune and MP2D guide pretrained diffusion models with multiple objectives at inference time (Tang et al., 2025; Kong et al., 2026). Against this background, ApexMO focuses on refinement within an explicit edit budget around each fixed original proposal, with recorded sequence ancestry. Our controlled evaluation examines how local-search operators and edit constraints affect predicted objective improvement and the number of selected representatives.

3 Method

Figure 1 summarizes the separation between proposal generation, fixed-anchor local search, and downstream candidate prioritization.

3.1 Problem formulation

Let $x \in \mathcal{A}^L$ be a peptide over the 20 canonical amino acids, and let a be its fixed generated anchor. Search sequences have lengths between 10 and 30. We minimize

$$\mathbf{f}(x) = \big(f_{\text{act}}(x), f_{\text{tox}}(x), f_{\text{hemo}}(x)\big), \qquad f_{\text{act}}(x) = \text{median}_{j\in\mathcal{J}} \log_{10}\left[\frac{\widehat{\text{MIC}}_j(x)}{1\,\mu\text{M}}\right], \tag{1}$$

where $\mathcal{J}$ indexes the 11 strains in Appendix C. Each $\widehat{\text{MIC}}_j(x)$ is the arithmetic mean of the 40 member predictions in linear MIC units, computed before taking the logarithm. The other two objectives are uncalibrated continuous scores for the toxic and hemolytic classes, respectively; lower scores are preferred. A peptide dominates another if all three scores are no larger and at least one is smaller.

Feasibility includes an anchor-relative edit bound

$$d_{\text{Lev}}(x, a) \le B(|a|), \qquad B(L) = \min\{\lfloor 0.25L \rfloor, 5\}, \tag{2}$$

as well as canonical residues, allowed length, task-specific applicability, activity uncertainty, and sequence novelty. Thus "five edits" denotes an absolute cap; shorter anchors can have smaller budgets. Feasibility gates restrict the search domain.

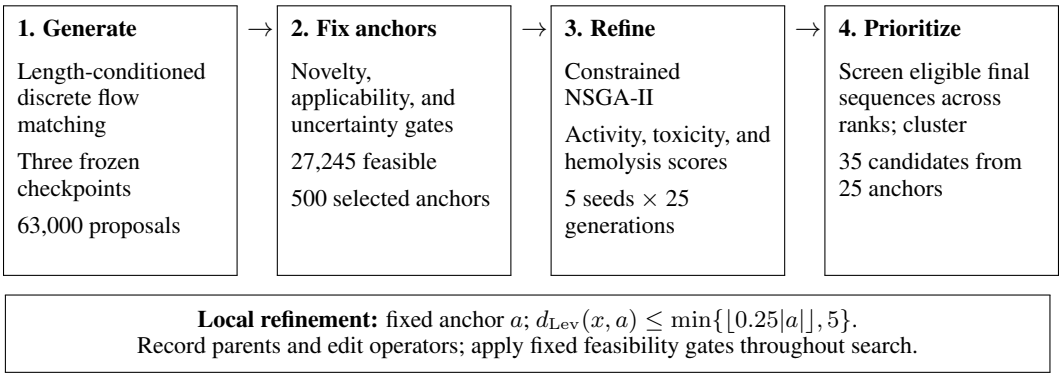


Figure 1: ApexMO workflow. Three independently trained, frozen checkpoints contribute 21,000 proposal events each (63,000 total). Selection fixes 500 anchors shared by the five main search runs. Each descendant retains its original anchor and ancestry. The cascade screens eligible sequences across Pareto ranks and selects representatives under 80%-identity/80%-coverage clustering. Counts refer to the main study; functional endpoints are predictions.

3.2 Length-conditioned generative prior

The generator is a bidirectional transformer conditioned on time and exact peptide length. Given a training sequence x_1 and an all-mask source x_0, the linear mixture path independently selects at each real residue position

$$x_t^i = \begin{cases} x_1^i, & \text{with probability } t, \\ \texttt{[MASK]}, & \text{with probability } 1 - t. \end{cases} \tag{3}$$

Padding remains fixed. The model predicts a distribution over the 20 amino acids, $p_\theta(x_1^i \mid x_t, t, L)$. Training samples $t \sim \mathcal{U}[0, 1 - \epsilon)$, with $\epsilon = 10^{-3}$, and minimizes token cross-entropy over all non-padding positions:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_1, t, x_t}\left[-\frac{1}{N_{\mathrm{tok}}} \sum_{b,i:\, i \leq L_b} \log p_\theta(x_{1,b}^i \mid x_{t,b}, t, L_b) \right]. \tag{4}$$

Here N_{tok} is the number of real residue tokens in the batch; the loss is not restricted to corrupted positions. Sampling uses exponential moving average (EMA) parameters, an all-mask initialization, and a discrete mixture Euler solver with step size $1/256$ up to $1 - \epsilon$. The posterior wrapper assigns zero target mass to the mask token. Architecture and training settings are given in Appendix B.

3.3 Predictive objectives and feasibility

We use the pretrained APEX ensemble (Wan et al., 2024a) with fixed weights. Its predictions for 11 selected strains define the median activity objective in Equation 1.

For toxicity and hemolysis, frozen ESM2-t30 embeddings (Lin et al., 2023) are averaged over residues, concatenated with sequence length, and passed to separate L2-regularized logistic-regression heads. Features are standardized using training data only. Labels originate from ToxinPred3 and HemoPI2 datasets (Rathore et al., 2024; 2025). Nested cluster-aware cross-validation selects the heads. Applicability domains constrain length and mean cosine distance to the five nearest task-specific training embeddings. Novelty gates exclude qualifying MMseqs2 matches at at least 50% identity and 80% bidirectional coverage to the training reference. The generator benchmark applies the same type of search-based criterion to a broader frozen reference union (Appendix A).

The measured-data stability classifier did not meet the prespecified deployment criterion, so stability was excluded from the Pareto objectives. Half-life predictions are retained only as exploratory annotations and sensitivity comparisons (Appendix C).

3.4 Anchored Pareto refinement

Preparation and feasibility filtering are followed by non-dominated sorting and diversity selection to obtain 500 anchors. Rank, crowding distance, and an 80%-identity/80%-coverage cluster cap select the initial population. Each anchor is retained as immutable metadata throughout its descendants' search history.

At each generation, constrained binary tournaments select parents by feasibility, Pareto rank, and crowding distance. Parent pairing seeks a distinct, equal-length mate with the same anchor and otherwise reuses the first parent. One-point crossover is attempted with probability 0.70; pairing a parent with itself leaves the sequence unchanged before mutation. Every offspring receives one mutation event: substitution, insertion, or deletion with probabilities 0.80, 0.10, and 0.10. Substitutions change one to three positions; replacement residues follow training-set amino-acid frequencies. Duplicate proposals and anchor-edit violations are rejected before scoring. Parent and offspring populations are combined, ranked, and reduced to 500 survivors with the diversity constraint. The main study uses five search seeds and 25 generations. Pseudocode and boundary handling are provided in Appendix D.

The lineage record identifies parents, operators, and the fixed anchor. A displayed primary-parent path is a summary of recorded ancestry; crossover can have two parents, and neither an alignment nor a primary-parent path attributes causal effects to individual edits.

4 Experimental protocol

Data and generator evaluation. The generator collection comprises 16,860 unique canonical AMPs of 5–64 residues from DBAASP and an additional laboratory AMP collection. MMseqs2 clustering at 67% identity and 80% bidirectional coverage yields cluster-disjoint training, validation, and test partitions of 13,478, 1,693, and 1,689 sequences. A separate 58,776-sequence Non-AMP collection is not used for generator training. Three independent training seeds are used. We evaluate sequence-order selectivity on 692 held-out AMPs with composition-preserving shuffles and exact-length Non-AMP controls; sampling audits use 50,000 generation events per model. The physicochemical distance-contrast analysis starts from 63,000 unfiltered events (1,000 per length from 10 to 30 per model) and restricts comparisons to shared lengths. Illustrative violins and amino-acid composition use separately defined subsets, detailed in Appendix B.

Generator comparison. Flow matching, AMP-Diffusion, and ARCADIAMP each make 10,500 attempts per sampling seed across three seeds, using one fixed checkpoint per system. The flow-matching checkpoint is from training seed 123; this benchmark differs from the three-checkpoint proposal library. Invalid, duplicate, and non-novel attempts remain in the denominator. Qualifying samples are unique within model/seed/length, exclude exact reference matches and qualifying MMseqs2 hits at 50% identity and 80% bidirectional coverage, and satisfy activity and uncertainty thresholds. Generation-only throughput is measured on NVIDIA H100 PCIe GPUs, excluding model loading and scorer runtime.

Search comparisons. We compare crossover-enabled NSGA-II, mutation-only NSGA-II, and random local search using the same 500 initial anchors and five paired seeds. Each method evaluates 500 initial anchors and 500 offspring in each of 25 generations, giving 13,000 unique sequence evaluations within each run, including scored offspring that fail subsequent feasibility checks. Early stopping is disabled in these comparisons. Random local search uses uniform parent selection and random feasibility-first survival; mutation-only NSGA-II retains tournament selection and Pareto survival. This control changes selection and survival jointly; evaluating all discovered feasible solutions distinguishes discovery from final-population retention.

Metrics. Final-population hypervolume uses the feasible sequences retained at the end of a run. Cumulative hypervolume uses all feasible sequences evaluated during that run, including initial anchors and sequences later discarded. Both are computed after the run from saved scores. The same feasibility definition and normalization are used within each comparison. Objective values are normalized by fixed generation-zero minima and ranges; the reference is $(1.05, 1.05, 1.05)$ in minimization space. Main-study and control values use their respective initial populations and are not

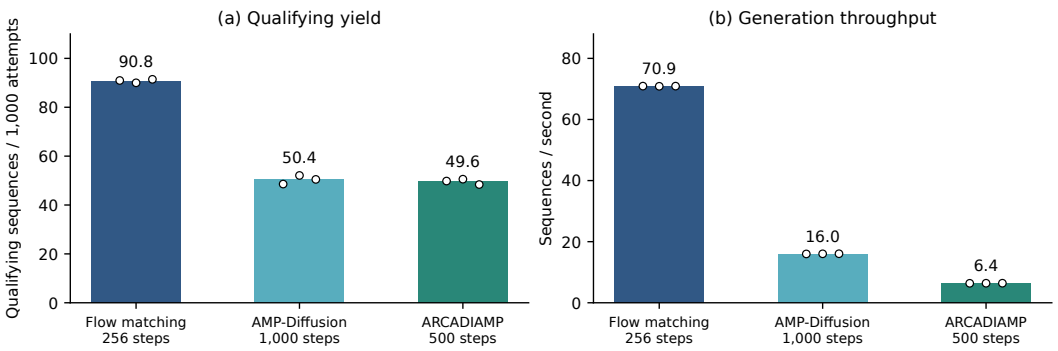


Figure 2: Generator comparison under matched attempt budgets. (a) Novel, predicted-active sequences recovered per 1,000 raw attempts. (b) Generation-only throughput on NVIDIA H100 PCIe GPUs, excluding model loading and scoring. Bars show means; open circles show all three sampling seeds, each with 10,500 attempts from one fixed checkpoint per system. Sampling steps appear below model names. The comparison is specific to the tested checkpoints and sampling configurations.

directly comparable. The same frozen screening cascade provides representative counts and anchor coverage. We report search repeats at the seed level; pooled sequence counts are not independent experimental replicates. Candidate–anchor distances are computed in the original ESM2 embedding space. All reported activity, safety, and half-life endpoints are predictions unless explicitly identified as measured training labels.

5 Results

5.1 The prior captures AMP-associated order and produces novel proposals

All three generator runs selected step 3,499 by minimum validation cross-entropy. Held-out test cross-entropy was 1.246 ± 0.003 (mean $\pm$ s.d.). In the matched sequence analysis, the ratios of masked-input pseudo-perplexity relative to AMPs were 1.102 (95% CI, 1.092–1.113) for composition-preserving shuffles and 1.402 (1.376–1.428) for Non-AMPs. The shuffled control supports an order-dependent signal beyond amino-acid composition; the denoising score is not an exact sequence likelihood.

Across 50,000 generation events per seed, every sequence used canonical residues and matched its requested length. Mean uniqueness and absence of exact training matches were both 99.7%; 91.0% and 46.8% of generation events had no qualifying MMseqs2 training-reference hit at the 80% and 50% identity thresholds, respectively, under the specified coverage and search settings. Generated distributions were closer to AMPs than Non-AMPs for all 12 assessed physicochemical properties, with bootstrap intervals excluding zero for 11. These analyses describe sequence resemblance and novelty, not measured function.

Under the matched attempt budget, flow matching yielded 90.8 qualifying sequences per 1,000 attempts, compared with 50.4 for AMP-Diffusion and 49.6 for ARCADIAMP (Figure 2). The throughput advantage is conditional on the tested checkpoints and their 256-, 1,000-, and 500-step sampling configurations. The comparison does not include training cost or end-to-end scoring cost.

5.2 Predictor performance defines the scope of optimization

Generated, AMP, and Non-AMP groups were length-matched at $n = 1,896$ each. Generated peptides had lower median APEX-predicted MIC than Non-AMPs for all 11 outputs, with ratios of 0.463–0.823. For each output, we used the absolute difference in median $\log_{10}$ MIC between Generated and Non-AMP sequences as the statistic for a two-sided, length-stratified permutation test (5,000 permutations). All 11 comparisons yielded $P \simeq 2.0 \times 10^{-4}$ after Benjamini–Hochberg correction across outputs. Generated peptides had higher median predicted MIC than known AMPs,

Table 1: Equal-budget local-search controls, using a separate frozen anchor population. All-survivor hypervolume is the median across five paired search seeds. Per-seed values appear in Appendix G. Representatives are cluster representatives before DBAASP prediction; counts and represented anchors are pooled across final populations.

Method	All-survivor HV	Representatives	Anchors represented
NSGA-II with crossover	1.092	48	29
Mutation-only NSGA-II	1.099	59	29
Random local search	0.804	7	7

with ratios of 1.05–1.31. The prior therefore occupied an intermediate predicted activity range (Figure S4a).

Mean out-of-fold area under the precision–recall curve (AUPRC; computed as average precision) was 0.922 for toxicity and 0.698 for hemolysis. On the complete frozen test sets, AUPRC was 0.938 and 0.830; sensitivity was 0.964 and 0.927, but specificity was 0.524 and 0.338. Removing test sequences homologous to development sequences produced smaller evaluation subsets and AUPRC values of 0.820 and 0.711. Dehomologized sensitivity was 0.922 and 0.866, at positive prevalences of 29.7% and 43.8% (complete tests: 50.0% and 46.4%). Classification thresholds were selected to target 95% sensitivity on out-of-fold development predictions and fixed for test evaluation (Appendix C).

Using the 11-strain APEX search scores, the uncertainty threshold retained 56,001 of 59,422 prepared generated sequences (94.2%). Toxicity and hemolysis applicability gates retained 80.4% and 87.6%, respectively. The reported anchor and final-variant sets passed these gates; the implementation can retain infeasible survivors if feasible sequences do not fill the population. The measured one-hour stability classifier achieved only 0.654–0.687 area under the receiver operating characteristic curve (AUROC) across three seeds and 105 unique sequences, motivating its exclusion from the Pareto objectives.

5.3 Pareto refinement improves the predicted final-population frontier

The five main runs produced 2,493 unique generation-25 variants, including 318 sequences on the pooled non-dominated frontier. All remained within their anchor-relative edit limits. All-survivor hypervolume increased from 0.873586 to 1.107933–1.117235, a 26.8–27.9% increase. Positive improvement denotes the initial minus final population median. Raw reductions were 0.060–0.257 $\log_{10}$ MIC (1.15–1.81-fold on the exponentiated median scale), 0.064–0.093 toxicity-score units, and 0.006–0.057 hemolysis-score units across seeds. Median predicted activity, toxicity, and hemolysis improved by 0.097–0.414, 0.254–0.368, and 0.028–0.254 initial-population interquartile ranges (IQRs), respectively (Figure 3).

Applying the same screening cascade to the 27,245 feasible pre-optimization sequences, the 500 selected anchors, and the 2,493 optimized variants retained 2, 2, and 52 representatives under 80%-identity/80%-coverage clustering, respectively. This comparison shows enrichment of the final population under the frozen cascade; it is not a comparison with equal-size random subsamples.

The separate equal-budget experiment yielded median final all-survivor hypervolumes of 1.092 for crossover-enabled NSGA-II, 1.099 for mutation-only NSGA-II, and 0.804 for random local search (Table 1). The common cascade, without a half-life gate, retained 48, 59, and 7 pooled representatives. These results descriptively support Pareto-based selection and survival under the tested local constraints. Adding crossover did not improve the median final all-survivor hypervolume relative to mutation-only NSGA-II under these constraints and evaluation budgets. Using all evaluated feasible solutions, median hypervolumes were 1.101, 1.106, and 0.952, respectively. Random search therefore discovered better points than its final population retained, while both Pareto methods retained a descriptive advantage under this common definition.

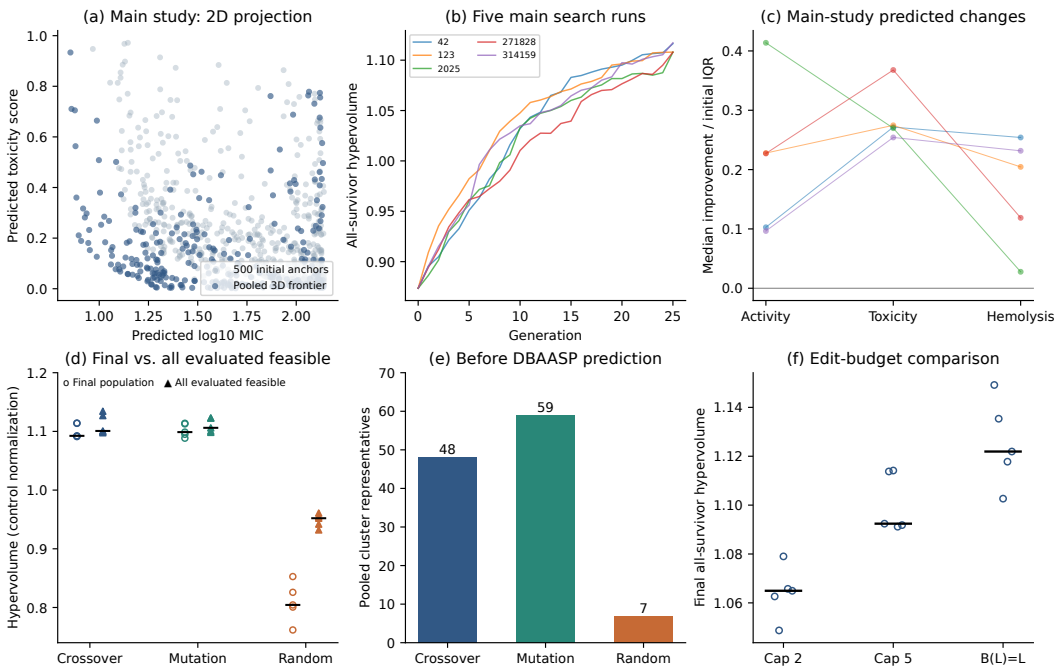


Figure 3: Predicted objective improvement and local-search controls. (a) Two-dimensional projection of initial anchors and the pooled three-objective frontier. (b) All-survivor hypervolume, one trajectory per main search seed. (c) Initial-minus-final median changes divided by the initial IQR, one line per seed. (d) Five paired seeds per control: open circles show final feasible populations, triangles show all evaluated feasible solutions, and black bars show medians. Each comparison uses fixed initial minima/ranges and reference $(1.05, 1.05, 1.05)$; main and control normalizations differ. (e) Pooled cluster representatives before DBAASP prediction. (f) Edit budgets: open circles are seeds and bars are medians. All objectives are predictions.

5.4 Edit budgets trade objective improvement against anchor proximity

In the equal-budget ablation, raising the absolute edit cap from two to five increased median final hypervolume from 1.065 to 1.092. A relaxed anchor-length bound, $B(L) = L$, increased it to 1.122. Median per-seed representative counts without the half-life gate were 10, 11, and 16. Pooled median candidate–anchor ESM2 cosine distances increased from 0.0087 to 0.0158 and 0.0267. Thus, tighter constraints retained closer sequences while limiting predicted improvement and the number of selected representatives. The main bound, $B(L) = \min\{\lfloor 0.25L \rfloor, 5\}$, is an intermediate operating choice rather than an established optimum.

The relaxed condition retains anchor metadata and same-anchor crossover. It is not mathematically equivalent to removing the edit constraint when length-changing mutations are permitted. Also, the two-edit condition required up to 5,000 unscored proposals per offspring, compared with the usual 100-attempt safeguard. Scored-sequence budgets were matched; proposal-generation effort was not.

5.5 The cascade yields traceable candidates whose membership is threshold-dependent

Of the 52 cluster representatives obtained without a predicted-half-life gate, 35 were classified as AMPs by the DBAASP Linear AMP Prediction tool and prioritized as candidates, covering 25 generated anchors. They differed from their anchors by three to five Levenshtein edits (Figure 4). Thirteen had lower values for all three predicted objectives; the remaining 22 represented trade-offs. Toxicity scores decreased for 33 of 35 candidates, hemolysis scores for 30, and both safety scores for 28. Internal-APEX median predicted MIC was 31.7 μM for 35 matched anchor instances and 26.2 μM for the candidates; 19 pairs improved. Median paired initial-minus-final changes were

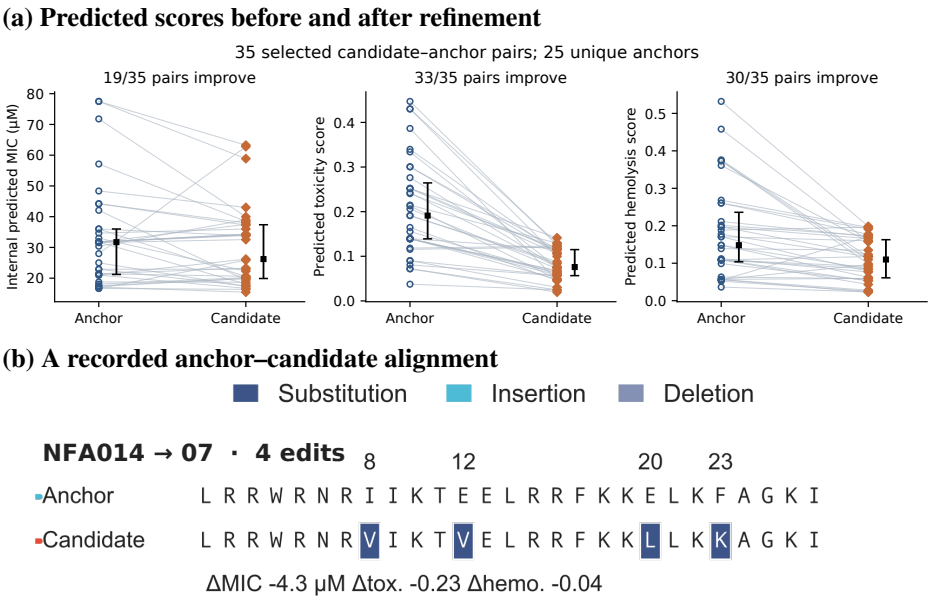


Figure 4: Candidate outcomes and traceability. (a) Paired predicted scores for 35 candidate–anchor pairs from 25 unique anchors; medians and IQRs weight repeated anchors once per candidate. Toxicity and hemolysis axes are predicted model scores; thresholds are frozen. Open blue circles denote anchors and orange diamonds denote candidates. Toxicity decreases for 33/35 candidates and hemolysis for 30/35. No significance tests are shown because selection uses these scores and candidates share anchors. (b) One recorded four-edit alignment (NFAMP00007), with predicted score changes; substitutions are highlighted. Full candidate profiles are shown in Figure S5. All endpoints are predictions.

0.008 $\log_{10}$ MIC, 0.107 toxicity-score units, and 0.040 hemolysis-score units. Repeated anchors are weighted once per candidate.

All 35 candidates had downstream APEX-pathogen predicted MIC below the computational screening cutoff of 128 μM (not a clinical susceptibility breakpoint) against four targets: *A. baumannii*, *E. coli* AIC221, *E. coli* AIC222, and vancomycin-resistant *E. faecium*. None did so against *K. pneumoniae*, *S. aureus* ATCC 12600, MRSA, or vancomycin-resistant *E. faecalis*. The aggregate activity objective therefore obscures target-specific weaknesses. Median candidate–anchor cosine distance was 0.017; the assigned anchor was the nearest among the 25 represented anchors for 24 candidates.

Candidate membership depended on screening choices. Without a half-life gate, tightening the APEX-pathogen MIC cutoff from 128 to 32 μM reduced representatives from 52 to 3; relaxing it to 256 μM produced 114. The primary 35-candidate set includes 18 of the 20 candidates selected with a one-hour half-life gate and 17 additional candidates. The other two candidates pass the no-gate criteria but are not selected as cluster representatives. All 52 primary representatives received DBAASP predictions using the same default tool and unmodified sequences. Other sensitivity analyses report cluster-representative counts before DBAASP classification. The one-hour-gated set serves as a secondary sensitivity comparator (Figure S9).

6 Limitations and Conclusion

ApexMO combines generative exploration with local multi-objective refinement around fixed sequence anchors. Its evidence is computational: the same predictors define the optimization landscape and much of the evaluation. Additional predictor-agreement filters share data sources with the primary safety models and do not provide independent biological validation. Applicability and uncertainty gates restrict extrapolation but do not eliminate surrogate exploitation. Search runs reuse a pooled proposal library and fixed selected anchors. Full-workflow robustness would require re-

peating generator training or checkpoint selection, proposal generation, anchor selection, and local optimization.

The generator benchmark compares released systems with different training histories and sampling schedules. Our optimizer comparisons are limited to the tested local-search controls; matched comparisons with existing sequence optimizers remain future work. Half-life predictions are exploratory annotations and do not determine primary selection. Synthesis, serum interactions, protease resistance, mechanism, and in vivo pharmacology were not optimized.

Within these limits, hypervolume for final populations and all evaluated feasible solutions favors anchor-constrained Pareto selection and survival over the tested random local-search control. Mutation-only NSGA-II has a slightly higher median than the crossover-enabled variant. The edit-budget analysis describes the trade-off between predicted improvement and sequence proximity. Prospective evaluation should compare candidates directly with their anchors across strain-resolved activity, cytotoxicity, hemolysis, and stability. Such measurements are needed to determine whether the recorded local refinements translate into biological improvement.

REPRODUCIBILITY STATEMENT

We provide the code, data splits, saved predictions, search records for each seed, and analysis scripts in the supplementary materials. We verified the reported search metrics by recomputing them from the saved predictions and records. Experimental procedures are described in Appendices A–E. Appendix F describes the software dependencies and the limits of these reproducibility checks.

ETHICS STATEMENT

This study focuses on the computational design of antimicrobial peptides. We plan to validate the selected candidates experimentally.

AI USE STATEMENT

Generative AI assisted with manuscript formatting, figure and reference checks, and verification scripts.

REFERENCES

K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6:182–197, 2002. doi: 10.1109/4235.996017.

D. Eisenberg, R. M. Weiss, and T. C. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences*, 81:140–144, 1984. doi: 10.1073/pnas.81.1.140.

I. Gat, T. Remez, N. Shaul, F. Kreuk, R. T. Q. Chen, G. Synnaeve, Y. Adi, and Y. Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133345–133385, 2024. doi: 10.52202/079017-4239.

H. Hu, C. Zhang, Z. Xu, J. Guo, A. Su, C. Li, and H. Duan. PepMSND: Integrating multi-level feature engineering and comprehensive databases to enhance in vitro/in vivo peptide blood stability prediction. *Digital Discovery*, 4:2478–2490, 2025. doi: 10.1039/D5DD00118H.

Z. Kong, Y. Dong, Y. Wu, Z. Liang, J. Wu, and H. Xu. MP2D: Constrained Monte Carlo tree-guided diffusion for multi-objective protein sequence design. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 5540–5548, 2026. doi: 10.24963/ijcai.2026/617.

B. P. Lazzaro, M. Zasloff, and J. Rolff. Antimicrobial peptides: Application informed by evolution. *Science*, 368(6490):eaau5480, 2020. doi: 10.1126/science.aau5480.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379:1123–1130, 2023. doi: 10.1126/science.ade2574.

J. Luo, K. Ding, and Y. Luo. Pareto-optimal sampling for multi-objective protein sequence design. *iScience*, 28(3):112119, 2025. doi: 10.1016/j.isci.2025.112119.

Y. Ma, Z. Guo, B. Xia, Y. Zhang, X. Liu, Y. Yu, N. Tang, X. Tong, M. Wang, X. Ye, J. Feng, Y. Chen, and J. Wang. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nature Biotechnology*, 40:921–931, 2022. doi: 10.1038/s41587-022-01226-0.

K. Markakis, S. Kim, C.-E. Tan, and I. Tagkopoulos. Discovery of potent low-toxicity antimicrobial peptides through diffusion modeling. *Nature Communications*, 17:8396, 2026. doi: 10.1038/s41467-026-75030-8.

L. McInnes, J. Healy, N. Saul, and L. Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3:861, 2018. doi: 10.21105/joss.00861.

N. G. Oliveira Júnior, C. M. Souza, D. F. Buccini, M. H. Cardoso, and O. L. Franco. Antimicrobial peptides: Structure, functions and translational applications. *Nature Reviews Microbiology*, 23: 687–700, 2025. doi: 10.1038/s41579-025-01200-y.

M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt, and M. Tartakovsky. DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Research*, 49:D288–D297, 2021. doi: 10.1093/nar/gkaa991.

A. S. Rathore, S. Choudhury, A. Arora, P. Tijare, and G. P. S. Raghava. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *Computers in Biology and Medicine*, 179:108926, 2024. doi: 10.1016/j.compbiomed.2024.108926.

A. S. Rathore, N. Kumar, S. Choudhury, N. K. Mehta, and G. P. S. Raghava. Prediction of hemolytic peptides and their hemolytic concentration. *Communications Biology*, 8:176, 2025. doi: 10.1038/s42003-025-07615-w.

S. Stanton, W. Maddox, N. Gruver, P. Maffettone, E. Delaney, P. Greenside, and A. G. Wilson. Accelerating Bayesian optimization for biological sequence design with denoising autoencoders. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 20459–20478, 2022. URL https://proceedings.mlr.press/v162/stanton22a.html.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017. doi: 10.1038/nbt.3988.

P. Szymczak, M. Możejko, T. Grzegorzek, R. Jurczak, M. Bauer, D. Neubauer, K. Sikora, M. Michalski, J. Sroka, P. Setny, W. Kamysz, and E. Szczurek. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nature Communications*, 14:1453, 2023. doi: 10.1038/s41467-023-36994-z.

S. Tang, Y. Zhang, and P. Chatterjee. PepTune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 59017–59065, 2025. URL https://proceedings.mlr.press/v267/tang25n.html.

M. D. T. Torres, L. T. Chen, F. Wan, P. Chatterjee, and C. de la Fuente-Nunez. Generative latent diffusion language modeling yields anti-infective synthetic peptides. *Cell Biomaterials*, 1:100183, 2025. doi: 10.1016/j.celbio.2025.100183.

M. D. T. Torres, Y. Zeng, F. Wan, N. Maus, J. Gardner, and C. de la Fuente-Nunez. A generative artificial intelligence approach for peptide antibiotic optimization. *Nature Machine Intelligence*, 8:841–856, 2026. doi: 10.1038/s42256-026-01237-5.

F. Wan, M. D. T. Torres, J. Peng, and C. de la Fuente-Nunez. Deep-learning-enabled antibiotic discovery through molecular de-extinction. *Nature Biomedical Engineering*, 8:854–871, 2024a. doi: 10.1038/s41551-024-01201-x.

F. Wan, F. Wong, J. J. Collins, and C. de la Fuente-Nunez. Machine learning for antimicrobial peptide identification and design. *Nature Reviews Bioengineering*, 2:392–407, 2024b. doi: 10.1038/s44222-024-00152-x.

B. Wang, P. Lin, Y. Zhong, X. Tan, Y. Shen, Y. Huang, K. Jin, Y. Zhang, Y. Zhan, D. Shen, M. Wang, Z. Yu, and Y. Wu. Explainable deep learning and virtual evolution identifies antimicrobial peptides with activity against multidrug-resistant human pathogens. *Nature Microbiology*, 10:332–347, 2025. doi: 10.1038/s41564-024-01907-3.

A DATASETS AND PARTITIONING

The generator dataset contains 16,860 unique canonical AMPs of 5–64 residues from DBAASP (Pirtskhalava et al., 2021) and an additional laboratory collection. For evaluation, we used 58,776 sequences labeled Non-AMP in the c_AMPs-prediction resource associated with Ma et al. (2022). Separate toxicity, hemolysis, and stability datasets were used to develop and evaluate the corresponding predictors. Table S1 summarizes the dataset and evaluation-set sizes.

Table S1: Dataset and evaluation-set sizes. Generator partitions are disjoint at the clustering level used for their construction. Smaller matched analysis sets serve different evaluation purposes and are not additional training partitions.

Dataset or evaluation set	Sequences / records
Unique AMP generator collection	16,860
Generator training / validation / test	13,478 / 1,693 / 1,689
Non-AMP reference	58,776
Matched AMP selectivity analysis	692 per group
Matched activity analysis	1,896 per group
Toxicity development / complete test	8,825 / 2,206
Hemolysis development / complete test	1,540 / 386
Toxicity / hemolysis dehomologized test	1,161 / 153
Measured stability	105 unique sequences; 106 records
Known-AMP activity-threshold benchmark	16,740

Generator partitioning. MMseqs2 (Steinegger & Söding, 2017) clustering used 67% identity, 80% bidirectional coverage, coverage mode 0, and cluster mode 1. The saved partition contains 5,975 clusters; the largest has 1,432 sequences (8.49% of the collection). Complete clusters were assigned to training, validation, and test in an approximately 80:10:10 ratio with partition seed 20260717. The generator test set was excluded from checkpoint selection and model configuration.

Novelty searches. MMseqs2 searches disabled sequence masking and composition-bias correction, used 80% bidirectional coverage, coverage mode 0, $k = 5$, sensitivity 7.5, and an E-value threshold of 10^6. The implementation additionally specifies a zero minimum ungapped score, unspaced k-mers, and a maximum of 20,000 returned hits. The sampling audit reports thresholds of 50% and 80% identity to the training reference. Initial library preparation excludes qualifying known-AMP matches at at least 80% identity. The optimization novelty gate excludes qualifying training-reference matches at at least 50% identity. The generator benchmark excludes exact reference matches and qualifying MMseqs2 hits at at least 50% identity and 80% bidirectional coverage to a frozen union of the full AMP collection used in this study and the disclosed AMP-Diffusion and ARCADIAMP training references. Keeping canonical sequences of 1–50 residues and deduplicating yields 46,211 reference sequences. Absence of a qualifying hit is a search-based criterion, not an exhaustive global-alignment identity bound.

Scorer evaluation sets. Canonical residues and task-specific length rules were applied before assigning toxicity and hemolysis labels. Development sequences were clustered at 50% identity and 80% bidirectional coverage, with complete clusters confined to cross-validation folds. Complete frozen tests contain 2,206 toxicity and 386 hemolysis sequences. Additional dehomologized subsets remove test sequences with a development-set match at at least 50% identity and at least 80% coverage of both sequences. Stability data contain 106 condition-specific records for 105 unique sequences because one sequence occurs under two retained measurement conditions.

B GENERATOR IMPLEMENTATION AND EVALUATION

The generator is a length-conditioned bidirectional discrete denoiser based on discrete flow matching (Gat et al., 2024). It contains six transformer blocks, hidden size 256, eight attention heads, an MLP ratio of four, and dropout 0.1. Time and length embeddings have dimension 128; the combined conditioning dimension is 256. Attention uses rotary positional embeddings, and conditioning

modulates the transformer blocks. The model supports lengths up to 64 residues and predicts the 20 canonical amino acids. The source consists of mask tokens at real residue positions and fixed padding elsewhere.

Probability path and training objective. The implementation uses a discrete mixture probability path with a polynomial convex scheduler of exponent 1.0. This gives the linear masked-source mixture described in the main text. Training samples time uniformly below $1 - 10^{-3}$. The cross-entropy implementation excludes padding but includes both corrupted and already visible residue positions. The resulting score is a denoising cross-entropy over all non-padding positions, including visible residues; it is neither a normalized sequence likelihood nor a masked-position-only loss.

Table S2: Generator training and sampling settings.

Setting	Value
Training seeds	42, 123, 2025
Optimizer	AdamW
Learning rate; betas	3×10^{-4}; $(0.9, 0.95)$
Weight decay; gradient clipping	0.03; norm 1.0
Batch size; precision	256; bfloat16
Warm-up	1,000 steps
Learning-rate schedule	Cosine decay to 10% of initial rate
EMA decay	0.999
Validation frequency	Every 500 steps
Checkpoint selection	Minimum deterministic validation token CE
Selected checkpoint	Zero-indexed step 3,499 (update 3,500) for each seed
Path exponent; endpoint offset	1.0; $\epsilon = 10^{-3}$
Sampling	EMA weights; discrete Euler step size $1/256$

The training configuration sets a maximum of 50,000 steps and an early-stopping patience of ten validation checks. It uses sequence sampling with a 2,048-sequence sortish buffer. The batch provider is a deterministic function of seed and step. Validation uses fixed sequence-specific random draws at eight times: 0.0625, 0.1875, 0.3125, 0.4375, 0.5625, 0.6875, 0.8125, and 0.9375. Checkpoint selection is performed independently within each training run. Validation is performed every 500 updates; all three runs stopped after 8,500 updates.

Sampling. Samples are generated in batches grouped by exact requested length, beginning from all-mask sequences. The discrete mixture Euler solver integrates from zero to $1 - \epsilon$ with step size $1/256$. The posterior contains the 20 amino-acid probabilities and an appended mask probability fixed to zero. The solver uses 256 intervals and decodes the final sample by drawing every residue from the target posterior evaluated at $t = 255/256$, resolving all remaining masks. The study generates 1,000 peptides at each integer length from 10 to 30 per checkpoint, totaling 63,000 events. A separate audit draws 50,000 events per checkpoint following the held-out AMP length distribution.

Sequence selectivity and composition. The selectivity analysis uses 692 held-out AMPs of 5–50 residues, exact-length Non-AMP controls, and composition-preserving shuffles. Matching masking patterns are used within each sequence set at the eight evaluation times. The reported pseudo-perplexity is the exponential of mean token cross-entropy under masked inputs, averaging over all non-padding residue positions rather than masked positions alone. Within each model and sequence, token losses are summed across evaluation times and divided by the total non-padding token count. Paired control-minus-AMP cross-entropy differences are averaged across the three models and matched sequence sets, then exponentiated to obtain the reported ratios. The 2,000 bootstrap resamples draw matched sequence sets and exponentiate each resampled mean difference.

The 12 physicochemical properties are calculated with the DBAASP `pept_char` tool and the Eisenberg–Weiss hydrophobicity scale (Eisenberg et al., 1984). The distance-contrast analysis starts from 63,000 unfiltered generation events and gives equal weight to the ten lengths shared by all groups: 11, 12, 16, 17, 23, 24, 27, 28, 29, and 30. The illustrative violins instead use prepared-pool property records restricted to these lengths: 28,339 Generated, 4,982 AMP, and 3,667 Non-AMP

sequences. With random seed 20260719, 500 observations per group and length are sampled with replacement, giving 5,000 plotted observations per group. These plotted draws are not independent biological replicates. For each property, the standardized contrast is

$$C_W = \frac{W(\text{Generated}, \text{NonAMP}) - W(\text{Generated}, \text{AMP})}{\text{IQR}(\text{AMP})}, \tag{5}$$

where W is computed on equal-weight mixtures across shared lengths, rather than averaging within-length distances. The denominator is the IQR of the length-balanced AMP mixture. Positive values indicate greater similarity to AMPs. We estimate intervals using 500 hierarchical resamples of seeds and lengths, then sequences. Each reference group is sampled once per selected length. The original AMP IQR is held fixed across these resamples. The plotting implementation replaces a nonpositive IQR with the unweighted standard deviation of AMP observations. It has no additional near-zero regularization; sensitivity to small positive denominators remains a limitation. Amino-acid composition uses 29,986 Generated, 4,982 AMP, and 3,672 Non-AMP unique sequences in the same ten lengths. Frequencies are averaged equally over lengths. One-sided exact paired sign permutations enumerate all $2^{10} = 1,024$ sign configurations to test greater Non-AMP frequency than each reference; Benjamini–Hochberg correction covers 20 amino acids separately for each comparison. Sequence-pattern analyses use Jensen–Shannon divergences for 1-mers, 2-mers, and 3-mers, normalized by a training–training resampling baseline; these intervals use 2,000 hierarchical bootstrap resamples.

Relative to Non-AMPs, generated sequences contain 3.29-fold more lysine, 1.59-fold more arginine, and 2.70-fold more tryptophan. Methionine, serine, and threonine have the largest Non-AMP enrichments; adjusted P values for each of these three residues are 0.0028 relative to AMPs and 0.0024 relative to Generated peptides. These distributional comparisons do not establish experimentally measured activity.

C Activity, Safety, and Stability Scorers

Activity. The internal activity scorer is the 40-model, 34-output APEX ensemble (Wan et al., 2024a). The selected 11 targets are *E. coli* ATCC 11775, *P. aeruginosa* PAO1 and PA14, *S. aureus* ATCC 12600, *E. coli* AIC221 and AIC222, *K. pneumoniae* ATCC 13883, *A. baumannii* ATCC 19606, methicillin-resistant *S. aureus* ATCC BAA-1556, vancomycin-resistant *E. faecalis* ATCC 700802, and vancomycin-resistant *E. faecium* ATCC 700221. AIC221 and AIC222 are MG1655-derived *E. coli* strains; their names follow the nomenclature in Wan et al. (2024a).

For each target, the 40 ensemble outputs are exponentiated to linear MIC units and averaged, then transformed back to $\log_{10}$ units. Activity is the median of these 11 $\log_{10}$ mean-MIC predictions. Ensemble dispersion uses sample standard deviation (denominator 39) of the member $\log_{10}$ predictions, with no target-specific cap. Ensemble dispersion is not a calibrated error bound. The median per-target ensemble standard deviation is constrained to at most 0.2771194711. This is the 95th percentile of the 16,740-sequence known-AMP scoring benchmark. The reference overlaps the generator training, validation, and test partitions by 13,380, 1,682, and 1,678 sequences, respectively, and is not held out from the generator. Independence from APEX's own training data is not established by these comparisons. Recorded activity-score quantiles are 1.3818304, 1.8882911, and 2.1178219 at probabilities 0.1, 0.5, and 0.9.

Activity comparisons match Generated, AMP, and Non-AMP sequences within ten length strata, with 1,896 per group. Per-target MIC ratios and intervals use 2,000 length-stratified bootstrap samples. Two-sided comparisons use the absolute difference of median $\log_{10}$ MIC, with 5,000 within-length label permutations. The plus-one estimate is $(b+1)/5{,}001$, where b counts permuted statistics at least as extreme as observed. All 11 Generated-versus-Non-AMP comparisons have $b = 0$, giving $P = 1/5{,}001$; Benjamini–Hochberg correction within this family leaves these identical values unchanged. They are resolution-limited estimates, not evidence of finer precision.

Safety heads and applicability. ESM2-t30, the 150M-parameter UR50D model (Lin et al., 2023), is frozen. Final-layer residue embeddings exclude padding and special tokens before averaging. Peptide length is appended and features are standardized using training data only. Toxicity and hemolysis use L2-regularized logistic regression with five outer and three inner

stratified group folds, repeated for seeds 42, 123, and 2025. The regularization grid is $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. Hemolysis also compares unweighted and balanced class weights. Each inner fit includes its own feature standardization; class-weight candidates are selected within the inner folds. Inner selection maximizes positive-class AUPRC, resolving ties by smaller C, then unweighted fitting. Deployment uses the modal outer-fold setting from seed 42; frequency ties favor smaller C, then lexical class-weight order. Both deployed heads use $C = 0.01$ without class weights. Operating thresholds come from pooled outer-fold predictions for seed 42.

Operating thresholds maximize specificity subject to at least 95% sensitivity in pooled out-of-fold predictions. The uncalibrated continuous scores, rather than thresholded labels, are used for Pareto optimization. Toxicity applicability requires length 4–35 and mean cosine distance to the five nearest training embeddings no greater than 0.0351. Hemolysis requires length 6–39 and distance no greater than 0.0432. These distance cutoffs are the 95th percentiles of leave-one-out mean-five-neighbor cosine distances within each development set; the displayed values are rounded. Distances use mean-residue embeddings, separately from the standardized classifier features. Self-neighbors are excluded when setting the cutoffs, but identical-sequence matches are not excluded when evaluating queries.

Table S3: Safety-model evaluation on complete frozen tests and dehomologized subsets. AUPRC is average precision. Thresholds are unchanged between each full test and its subset. Values are rounded for display.

Task	Evaluation set	n	AUROC	AUPRC	Sensitivity	Specificity
Toxicity	Complete	2,206	0.9303	0.9384	0.9637	0.5245
Toxicity	Dehomologized	1,161	0.8879	0.8202	0.9217	0.5282
Hemolysis	Complete	386	0.8301	0.8298	0.9274	0.3382
Hemolysis	Dehomologized	153	0.7473	0.7112	0.8657	0.3488

The dehomologized evaluations reuse predictions and thresholds without retraining. They exclude 1,045 toxicity and 233 hemolysis sequences. Positive prevalence changes from 50.0% to 29.7% for toxicity and from 46.4% to 43.8% for hemolysis, affecting interpretation of AUPRC differences. The toxicity subset contains 318 true positives, 27 false negatives, 431 true negatives, and 385 false positives. The hemolysis subset contains 58, 9, 30, and 56, respectively. Deployment thresholds are 0.1420483136 for toxicity and 0.2063112305 for hemolysis. Limited specificity constrains candidate recovery, while the observed false negatives and reduced sensitivity on dehomologized subsets limit confidence in sequences receiving negative toxicity or hemolysis predictions.

Stability. Measured half-life labels are curated from PepMSND (Hu et al., 2025). The one-hour classifier uses ESM2 embeddings, length, species, and matrix. Its 106 condition-specific records comprise 37 positive and 69 negative records from 105 unique sequences. Each record receives inverse within-sequence multiplicity weight; the repeated sequence therefore contributes two half-weight records, and the total positive/negative weights are 37/68. Records share a sequence-cluster group across folds. ROC curves compare measured labels with out-of-fold predicted scores, not with refitted training predictions. Advancement requires AUROC at least 0.70 and AUPRC at least 0.15 above prevalence for every seed, together with fixed checks on precision, recall, baseline improvement, endpoint sensitivity, and cluster separation. Observed AUROCs of 0.687, 0.676, and 0.654 fail the AUROC criterion, so stability is not optimized. The primary analysis does not use a predicted-half-life filter; a one-hour filter is evaluated as a sensitivity condition. A separate frozen ridge regressor predicts $\log_{10}$ half-life in hours from standardized mean ESM2 residue embeddings, without species or matrix covariates; the classifier validation protocol does not apply to this regressor. For sensitivity seeds 123 and 2025, recorded Spearman correlations are 0.304 and 0.360. MAEs are 0.7896 and 0.7982 in $\log_{10}$ half-life units, versus 0.7098 for the training-median baseline (11.2% and 12.5% higher). The regressor therefore did not outperform this baseline in the recorded evaluations. Half-life predictions therefore serve only as exploratory annotations in the primary analysis. Without a half-life gate, ranking, clustering, and DBAASP prediction yield 52 representatives and 35 predicted-AMP candidates. The 20-candidate one-hour-gated set is retained as a sensitivity comparator; it does not establish experimental stability or selection robustness.

D SEARCH IMPLEMENTATION AND COMPARISON PROTOCOLS

Preparation applies residue/length validation, deduplication, exact-match removal, and known-AMP homology filtering, retaining 59,422 generated sequences. Applicability, uncertainty, and training-reference novelty requirements retain 27,245 for initial ranking. The objectives are internal APEX median activity, toxicity, and hemolysis. Non-dominated rank and crowding distance order candidates; MMseqs2 clustering at 80% identity and 80% bidirectional coverage retains at most five members per cluster until 500 anchors have been selected. Clustering uses coverage mode 0 and cluster mode 1.

Main search. Seeds are 42, 123, 2025, 271828, and 314159. Every offspring retains an original anchor and must satisfy $d_{\mathrm{Lev}}(x, a) \leq \min\{\lfloor 0.25|a| \rfloor, 5\}$. Mate selection excludes the first parent and requires equal length and a shared anchor. When no distinct mate exists, the first parent is supplied twice to the crossover function. The nominal 0.70 crossover draw and a uniformly sampled internal breakpoint can therefore be logged even though no recombination occurs. We distinguish nominal labels, distinct-parent crossover, and pre-mutation sequences different from both parents. The mutation probabilities are 0.80 for substitution and 0.10 each for insertion and deletion. The substitution count is uniform over one, two, or three; positions are drawn uniformly without replacement. Insertion positions are uniform over the gaps, including both termini; deletion positions are uniform over residues. Insertions at length 30 and deletions at length 10 are converted to substitutions, with both the original draw and executed operator retained as separate metadata fields. Residues are sampled from training frequencies; substitutions exclude the existing residue.

Algorithm S1: anchored local Pareto refinement

Input: generated anchor population P_0; frozen scorers; feasibility rules; population size $N = 500$; generations $G = 25$.

Initialize: evaluate P_0; assign each sequence itself as its fixed anchor; initialize sequence history and lineage records.

For $g = 1, \ldots, G$:

1. Select parents by constrained binary tournaments. Restrict a crossover mate to the same length and anchor.
2. Draw nominal one-point crossover with probability 0.70, followed by one mutation event. A missing mate is replaced by the first parent, so a crossover label alone does not establish recombination.
3. Reject forbidden or duplicate sequences and proposals violating length or anchor-edit bounds; retry up to the per-offspring limit. Exhaustion raises an error and stops the run rather than relaxing a rule or scoring a smaller batch.
4. Evaluate offspring with the frozen scorers and annotate applicability, uncertainty, and novelty feasibility.
5. Combine parents and offspring. Apply constrained Pareto sorting, crowding, and cluster-limited survival to retain N sequences.
6. Save population scores, sequence history, anchor metadata, and parent/operator records; log the search-time hypervolume from feasible survivors with stored rank zero. Reported hypervolumes are recomputed from all feasible survivors or all evaluated feasible solutions, as defined below.

Output: final population, evaluation history, and recorded ancestry.

Figure S1: Operational summary of the main search. The main experiment also checks early stopping with patience five and minimum hypervolume improvement 0.001; all five reported main runs completed 25 generations. Equal-budget controls disable early stopping.

Exact duplicates are rejected, including previously generated sequences tracked in the run history. The ordinary proposal safeguard permits at most 100 attempts per offspring; exceeding it raises an offspring-exhaustion error. Each generation evaluates 500 offspring. Ties in Pareto rank and crowding distance are broken deterministically using the sequence SHA-256 hash. Each seed writes generation-specific populations, lineage records, sequence-history files, random-state records, and

a scored-sequence cache. Clusters are recomputed on each combined population. If the cluster cap leaves fewer than 500 survivors, selection raises an error. Feasible records precede infeasible ones; infeasible records are ordered by a supplied aggregate violation if present, otherwise by sequence hash after an infinite-violation default. Infeasible survivors are permitted when needed to fill the population. These records support provenance and the retrospective cumulative-set analysis.

Hypervolume. Let m_k and r_k be the minimum and range of objective k at generation zero. Normalize $f_k(x)$ as $z_k(x) = (f_k(x) - m_k)/r_k$, replacing a zero range with one. Let R_g contain the feasible surviving records whose stored Pareto rank equals zero. Stored ranks are assigned in the combined parent–offspring selection pool and are not recomputed after survivor selection. The metric logged during search was

$$\mathrm{HV}_{\mathrm{stored}}(P_g) = \mathrm{Vol}\left(\bigcup_{x \in R_g} \prod_{k=1}^{3} [z_k(x), 1.05]\right), \tag{6}$$

retaining only finite points strictly below the reference in all coordinates. This definition can differ from hypervolume recomputed from all feasible survivors because ranks are not updated after survivor selection. It does not use the cumulative feasible set. For the reported final-population results, replace R_g by all feasible current survivors to obtain $\mathrm{HV}_{\mathrm{all}}(P_g)$. For the cumulative-set metric, replace it by all feasible scored sequences observed through generation g, including initial anchors. Accepted-offspring lineage records verify the anchor-edit gate before scoring. Neither definition relies on stored Pareto ranks. The normalization is fixed during the run. Values above one are possible because optimized objectives can improve beyond the generation-zero minima. Comparisons use the same normalization when sharing a frozen initial population.

Equal-budget controls. A separate frozen population of 500 anchors is shared by crossover-enabled NSGA-II, mutation-only NSGA-II, and random local search. All use the same mutation operators, feasibility rules, diversity constraints, and five paired seeds. Each run scores 500 anchors and 12,500 offspring, for 13,000 unique sequences within the run, including scored sequences that subsequently fail feasibility gates. The per-sequence cache retains scorer outputs and feasibility annotations; rejected unscored proposals do not consume this budget. This is not a matched wall-time or total-compute budget. Mutation-only disables crossover but retains tournament selection and Pareto survival. Random local search chooses parents uniformly, disables crossover, and uses random feasibility-first survival, so objectives do not determine survival within a feasibility group. Final populations are pooled and deduplicated within method, then passed through the common cascade. Main-study finalization preserves per-seed rank/crowding as provenance, but recomputes rank and crowding on the pooled externally eligible set using downstream APEX-pathogen activity, toxicity, and hemolysis. The retrospective evaluation also applies the same feasibility rules to all evaluated solutions within each run, without sharing discoveries between methods. Final-population and cumulative-set hypervolumes are reported separately.

Edit-budget ablation. The three reported bounds are $\min\{\lfloor 0.25L \rfloor, 2\}$, $\min\{\lfloor 0.25L \rfloor, 5\}$, and L, where L is anchor length. We refer to the last as the relaxed condition. It retains anchor metadata and same-anchor crossover; because sequence lengths can change, a bound of L is not guaranteed to be inactive. Figure panels label this condition as a relaxed anchor-length bound. The two-edit condition permits up to 5,000 unscored proposals per offspring. Every generation still scores 500 offspring, preserving the 13,000-evaluation budget but not matching proposal effort. Locality is assessed with edit distance, length-normalized edit distance, and ESM2 cosine distance. The exact anchor is also ranked by embedding distance among all 500 initial sequences.

Generator benchmark. Each model generates 500 peptides at each integer length from 10 to 30 for seeds 42, 123, and 2025, totaling 31,500 attempts per model, written in chunks of 100. Qualifying sequences are unique within model/seed/length, have neither an exact match nor a qualifying MMseqs2 hit to the frozen reference union defined in Appendix A, and satisfy an internal APEX activity score below 1.89 and median ensemble dispersion at most 0.28. These are the known-AMP median activity and 95th-percentile dispersion cutoffs. These activity thresholds differ from the 128 μM downstream candidate cutoff. The three benchmark seeds vary sampling from a single checkpoint per system. Flow matching uses the training-seed-123 EMA checkpoint (selected

by validation cross-entropy) and 256 Euler steps; AMP-Diffusion uses its released checkpoint and 1,000 steps; ARCADIAMP uses its released D3PM checkpoint and 500 steps (Torres et al., 2025; Markakis et al., 2026). Throughput includes generation only, excluding startup, model loading, and downstream scoring. No replacement sampling is used to compensate for failed attempts.

E CANDIDATE PRIORITIZATION, LINEAGE, AND SENSITIVITY

The five main generation-25 populations contribute 2,500 instances and 2,493 unique sequences. Eligibility is not restricted to the pooled non-dominated frontier. Candidates must satisfy the original edit, applicability, uncertainty, and novelty requirements and the internal thresholds in Table S4.

Table S4: Frozen candidate-screening thresholds. Safety scores are model-specific outputs, not measured toxicity or calibrated probabilities.

Screen	Criterion
Internal APEX median MIC	$< 128\ \mu$M
Internal toxicity score	< 0.1420483136
Internal hemolysis score	< 0.2063112305
Median activity uncertainty	≤ 0.2771194711 ($\log_{10}$ MIC s.d.)
Predicted half-life	Annotation only; no primary gate
APEX-pathogen median MIC	$< 128\ \mu$M
ToxinPred3 / HemoPI2 agreement thresholds	0.38 / 0.58
Representative clustering	80% identity, 80% bidirectional coverage
Final prioritization	DBAASP-predicted general-AMP label

After internal screening, candidates are rescored with the released eight-model APEX-pathogen implementation. ToxinPred3 and HemoPI2 provide secondary safety-agreement checks using the saved Non-Toxin and Non-Hemolytic labels at their operating thresholds. Rounded displayed scores are not re-thresholded; a displayed ToxinPred3 score of 0.38 can retain a Non-Toxin label. Passing sequences from all Pareto ranks are clustered at 80% identity and coverage. Pareto rank and crowding distance are recomputed on the pooled eligible set using downstream APEX-pathogen median $\log_{10}$ MIC, toxicity, and hemolysis. Rank, descending crowding distance, and sequence-hash tie-breaking select one representative per cluster. Duplicate sequences use the lowest source-seed number for the displayed anchor; all final-instance lineages are retained in the accompanying data. Unmodified canonical representatives are submitted to the DBAASP Linear AMP Prediction interface, using the default "Prediction of general antimicrobial activity" tool. Representatives predicted to be AMPs by this tool are retained for experimental prioritization; these computational labels do not establish antimicrobial activity. Selecting one representative before this filter means that rejecting it does not imply rejection of every eligible member of its cluster.

The cascade retains 832 sequences after the internal screen, 155 after the APEX-pathogen MIC screen, 103 after secondary safety agreement, 52 cluster representatives, and 35 DBAASP-predicted AMP candidates. The final set covers 35 clusters and 25 anchors. Candidate-level predicted profiles are displayed in Figure S5. The statistical analyses use these selected predictions and are descriptive.

Embedding analysis. Frozen ESM2 mean-residue embeddings are L2-normalized in 640 dimensions. For visualization, the archived UMAP (McInnes et al., 2018) coordinates of 15,000 Non-AMPs, 10,000 DBAASP AMPs, and 2,582 laboratory AMPs are held fixed. We reconstruct the UMAP neighbor model from their reference embeddings, restore those archived coordinates, and transform the 25 anchors and 35 candidates. UMAP v0.5.12 uses 30 neighbors, minimum distance 0.2, cosine metric, and seed 20260721. Queries do not fit or move reference coordinates. Exact distances, nearest-anchor assignments, and ratios to non-matching-anchor distances are computed in the original embedding space, not on the UMAP projection. The final-set nearest-anchor comparison concerns the 25 represented anchors; the edit-ablation anchor-rank analysis concerns all 500 starting sequences.

Lineage. Each displayed state is globally aligned to its initial anchor; insertion slots are expanded consistently across states. Stepwise plots follow the recorded primary-parent path. For crossover, the

primary parent is the one contributing more pre-mutation residues, with lexical tie-breaking. Candidate NFAMP00011 has a six-state displayed path comprising two crossover and three mutation-only transitions, with a final anchor distance of three edits. Its G24 label denotes creation in generation 24; the candidate survives into the generation-25 final population. A nominal crossover label need not establish recombination, as quantified in Appendix G. Alignments and ancestry displays do not establish causal effects of individual substitutions.

Threshold sensitivity. Sensitivity analyses use the 2,493 unique main-run variants with fixed scores. One-factor grids vary internal APEX and APEX-pathogen MIC thresholds over 32, 64, 128, and 256 μM; toxicity and hemolysis thresholds over 80%, 100%, and 120% of baseline; and uncertainty over the known-AMP 90th, 95th, and 99th percentiles. Stability is examined with no filter and half-life thresholds of 0.5, 1, and 2 hours. Targeted two-factor grids cross the two MIC thresholds, toxicity with hemolysis, and uncertainty with stability. All other criteria remain fixed; the primary baseline has no half-life gate.

Each condition is independently re-ranked using APEX-pathogen median $\log_{10}$ MIC, toxicity, and hemolysis, then clustered. Recorded endpoints include stage counts, rank distributions, representative and anchor counts, objective distributions, representative-set Jaccard similarity, and retention of current candidates. Re-clustering can change which sequence represents a cluster; loss of a current representative does not necessarily mean it failed a relaxed threshold. No DBAASP label is imputed for an alternative representative.

Tightening the APEX-pathogen cutoff to 32 μM yields three representatives and retains one primary candidate as a representative; relaxing to 256 μM yields 114 and retains 27. A two-hour half-life threshold yields 24 representatives and retains 15 primary candidates; a one-hour gate yields 31 and retains 18. The no-gate primary condition yields 52 and retains all 35. Jaccard similarities between non-identical half-life-filter representative sets range from 0.41 to 0.76. The primary baseline reproduces all 52 submitted representative sequences and their assigned anchors. DBAASP classifies 35 representatives as AMP and 17 as Non-AMP; all 29 representatives shared with the 31-representative one-hour-gated set have concordant labels. This agreement does not establish robustness of the 20-candidate one-hour-gated set. Both the no-gate and one-hour-gated conditions rank candidates using downstream APEX-pathogen activity after external screening; internal APEX defines the search objective and initial screening.

Additional aggregation and screening analyses. For the 35 primary candidates, pathogen-aggregation checks compute median, mean, 75th-percentile, and maximum MIC on the linear μM scale, and coverage below 32, 64, 128, and 256 μM. Rankings use ascending MIC or descending 128 μM coverage with minimum ranks for ties. Leave-one-pathogen-out summaries remove each target in turn; Spearman correlations are descriptive. The matched pre-optimization comparison applies identical feasibility, activity, safety, agreement, and clustering rules to the 27,245-sequence feasible pool, 500 anchors, and 2,493 variants, with and without the exploratory stability filter. It stops at the reproducible cluster-representative stage before DBAASP classification.

F Statistical interpretation and reproducibility scope

Generator training seeds, search seeds, sequence-level samples, and final candidate pairs represent different sources of variation. Bootstrap procedures retain the pairing or length structure of the corresponding comparison. Benjamini–Hochberg correction is applied within the stated comparison families. Candidate comparisons are descriptive because selection uses the same scores and 35 candidates share 25 anchors. We emphasize paired directions of change (19/35 for activity, 33/35 for toxicity, and 30/35 for hemolysis); machine-readable paired effects and saved-precision scores are supplied with the supplementary data. No hypothesis-test significance is claimed for the selected candidates. Seed-level tests on five optimization repeats are likewise interpreted descriptively.

The main-study workflow uses `configs/pipeline.yaml`: five search seeds, 25 generations, internal APEX activity, subsequent APEX-pathogen screening, and no stability gate. Primary candidate selection uses the saved representatives and author-obtained DBAASP labels. Configuration usage and the separate example workflow are documented in the code README.

The code snapshot specifies `flow-matching==1.0.10`, `torch==2.8.0`, and `scikit-learn==1.7.1`; the recorded software environment and model hashes are supplied. The supplementary archive includes saved scores, populations, and accepted-offspring lineage for all 30 runs, as well as offline analysis code. External scorer dependencies are documented separately. Reproducibility checks cover offline metric reconstruction from saved scores and search records; numerical reproducibility of a fresh end-to-end run has not been evaluated.

G Additional Experimental Results

We report per-seed hypervolumes and crossover statistics, together with precision–recall curves for the safety predictors.

Table S5: Final all-survivor hypervolume by paired search seed. Main and control normalizations differ.

Seed	Main	Crossover	Mutation	Random	Cap 2	Relaxed
42	1.116391	1.092409	1.088441	0.852650	1.062634	1.149160
123	1.107999	1.113775	1.095070	0.804348	1.048726	1.135353
2025	1.107938	1.114154	1.113830	0.825850	1.079018	1.102650
271828	1.107933	1.091172	1.112839	0.800665	1.065735	1.117770
314159	1.117235	1.091827	1.098812	0.761711	1.064950	1.121928

Table S6: Hypervolume over all evaluated feasible solutions, by search seed.

Seed	Main	Crossover	Mutation	Random	Cap 2	Relaxed
42	1.122412	1.097711	1.098373	0.932123	1.066278	1.152173
123	1.114328	1.134385	1.100586	0.952082	1.055508	1.141872
2025	1.116008	1.126919	1.122772	0.954337	1.084127	1.122736
271828	1.115449	1.099512	1.122830	0.960938	1.069395	1.128318
314159	1.125993	1.100766	1.106085	0.942082	1.074632	1.127442

Crossover exposure. Across the five main runs, 44,033 of 62,500 accepted offspring carry a nominal crossover label, but 9,239 of those labels use the same parent twice. Distinct-parent crossover accounts for 34,794 offspring (55.7%), of which 7,413 survive their creation generation. Only 14,218 offspring (22.7%) have a pre-mutation recombinant different from both parents; 2,821 of these survive. Initially, no parent has a distinct compatible same-anchor mate.

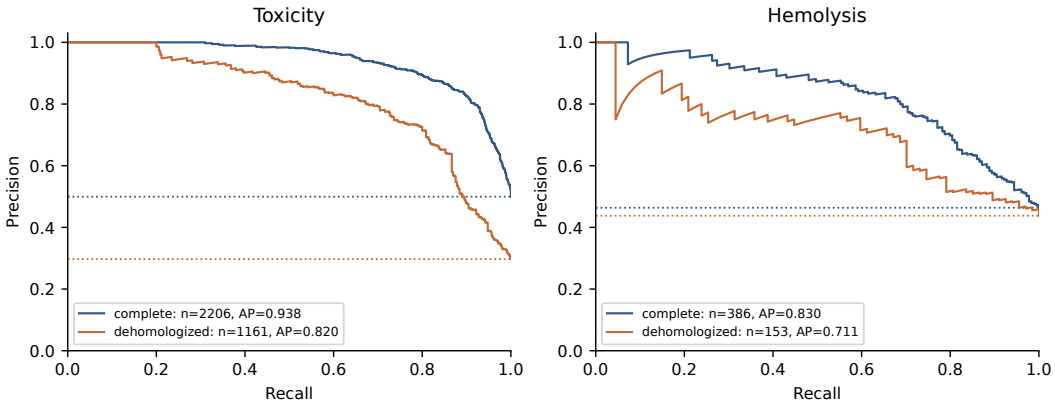


Figure S2: Complete frozen-test and dehomologized precision–recall curves using the original scores and unchanged operating thresholds. AP denotes average precision. Dotted lines are the respective positive prevalences; sample counts are shown in each legend.

H SUPPLEMENTARY DATA AND CODE INDEX

The supplementary ZIP contains machine-readable CSV/TSV tables, compressed saved search records, model-evaluation files, and analysis code. The archive includes internal activity and safety scores, DBAASP labels for all 52 representatives, and candidate/anchor ESM2 embeddings. File checksums and export provenance are supplied with the archive.

Table S7: Machine-readable supplementary materials. D identifiers refer to directories in the archive; these are separate from the numbered tables embedded in this PDF.

Directory	Contents
D01	Dataset inventory, generator splits, cluster assignments and partition audit.
D02	Generator evaluation, sampling repeats, length-level summaries and saved distributions.
D03	Scorer labels, out-of-fold and held-out predictions, and performance records.
D04	All 30 search runs: evaluated scores, 25 populations per run, lineage, normalization, hypervolume and paired controls.
D05	Matched screening controls, edit-budget comparisons and representative counts.
D06	Per-generation crossover exposure and repeated-sequence ancestry.
D07	The 35 primary candidates, 25 anchors, saved scores, DBAASP labels, lineage and one-hour-gated 20-candidate comparison.
D08	Threshold and half-life sensitivity, with eligibility separated from representative retention.
D09	Strain-resolved predictions, aggregation, coverage and leave-one-target-out summaries.
D10	Recorded training settings, software environment, model hashes and export provenance.
D11	Source data and metadata for supplementary figure panels.

I SUPPLEMENTARY FIGURES

Figures S3–S9 provide additional distributions, predictor diagnostics, candidate profiles, and sensitivity analyses.

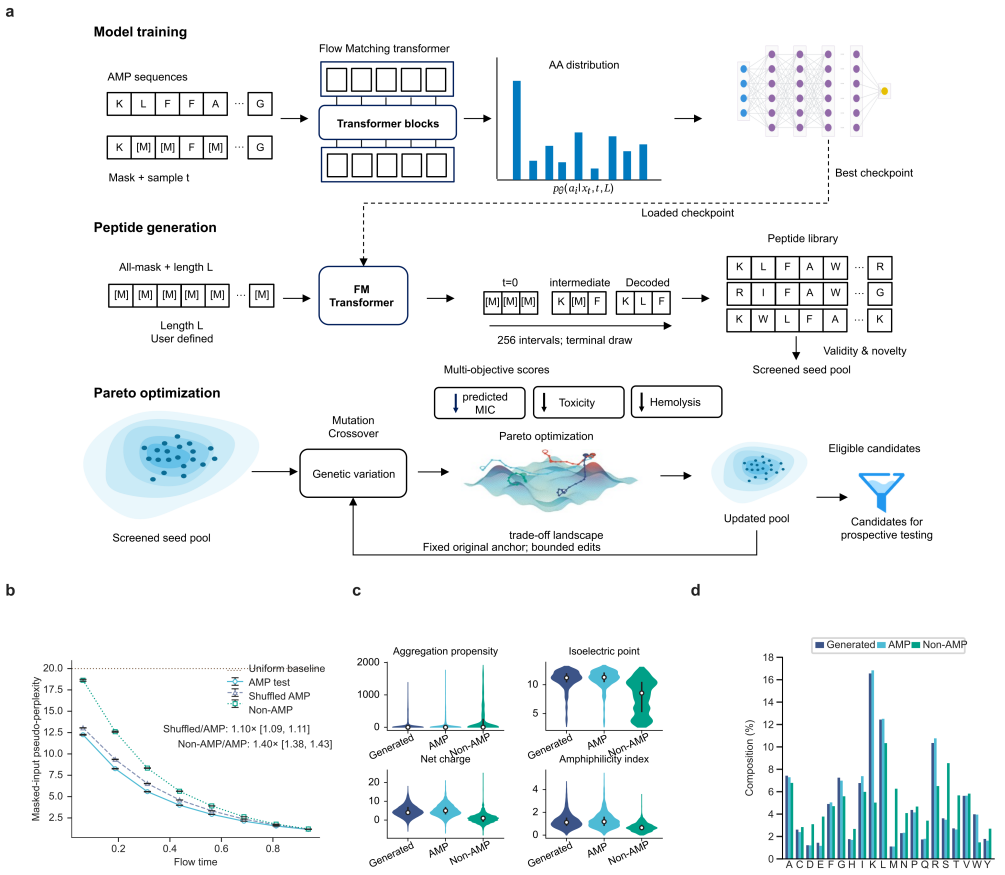


Figure S3: Generator and workflow analyses. (a) Flow-matching training, sampling, and local optimization. The schematic labels the separately decoded target sample explicitly; the solver grid ends at $1 - \epsilon$ and the final posterior draw is evaluated at $255/256$. Local descendants remain constrained to their fixed original anchor by Equation 2. Final screening includes eligible sequences from all Pareto ranks, as specified in Appendix E. (b) Mean masked-input token pseudo-perplexity for 692 matched sequence sets across three models; error bars show s.d., and ratios use paired-bootstrap confidence intervals. (c) Four illustrative physicochemical distributions, with 500 replacement draws per shared length and group (5,000 plotted points per group); white points and black lines indicate plotted-sample medians and IQRs. (d) Length-balanced amino-acid composition with one-sided exact paired sign-permutation tests for Non-AMP enrichment and Benjamini–Hochberg correction across 20 amino acids per comparison.

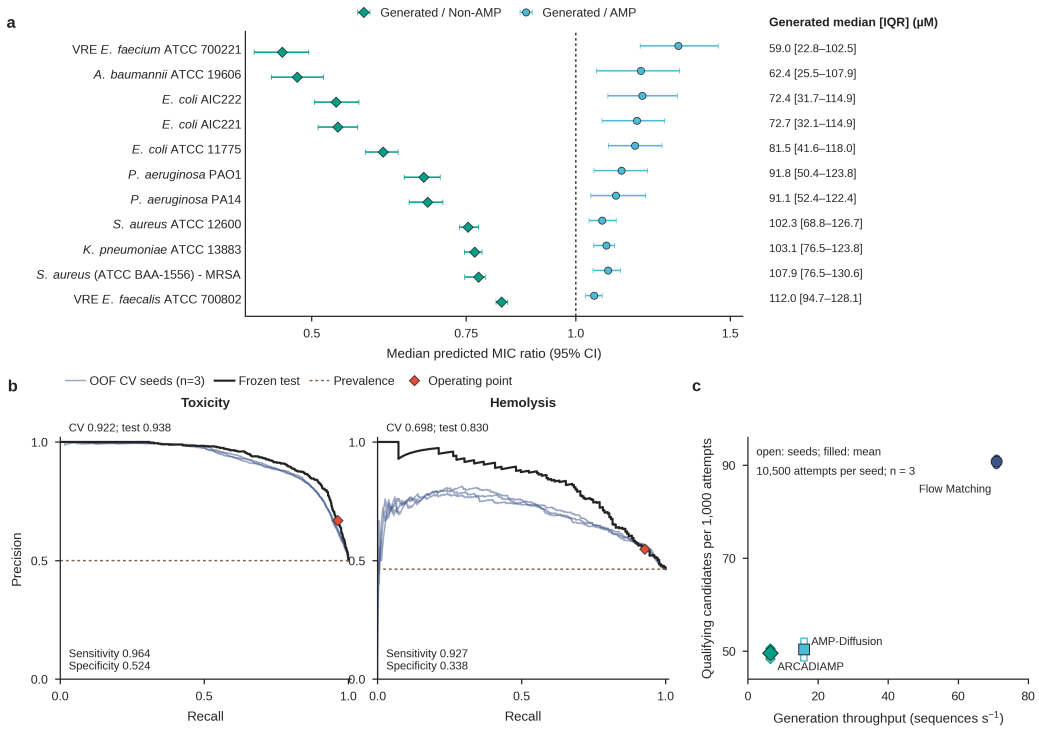


Figure S4: Activity, safety, and generator comparisons. (a) Ratios of median APEX-predicted MIC for 1,896 length-matched generated sequences versus AMP and Non-AMP references across 11 outputs; error bars are bootstrap 95% intervals. The right column reports generated medians and IQRs. (b) Safety precision–recall curves for three out-of-fold seeds and complete frozen tests; AUPRC denotes average precision. Dehomologized counts, prevalences, and performance appear in Table S3 and its accompanying text; a companion plot shows both curve sets in Figure S2. Diamonds mark operating points and dotted lines show prevalence. Development/test sizes are 8,825/2,206 for toxicity and 1,540/386 for hemolysis. (c) Generation-only throughput versus qualifying candidates per 1,000 attempts. Open points are three seeds; filled points are means. Each seed uses 10,500 attempts.

ID	Anchor	Edits	Half-life (h)
01	NFA005	5	0.15
02	NFA008	4	5.19
03	NFA018	3	0.41
04	NFA012	5	7.18
05	NFA023	4	3.55
06	NFA004	5	1.81
07	NFA014	4	3.26
08	NFA005	5	0.24
09	NFA008	5	0.23
10	NFA005	5	0.31
11	NFA018	3	1.48
12	NFA015	5	5.72
13	NFA025	5	0.01
14	NFA003	5	0.64
15	NFA010	5	0.18
16	NFA009	5	0.02
17	NFA002	5	13.55
18	NFA023	5	10.44
19	NFA006	5	3.31
20	NFA015	5	1.84
21	NFA005	5	0.10
22	NFA019	5	7.65
23	NFA012	4	1.65
24	NFA016	5	4.98
25	NFA024	5	4.63
26	NFA020	5	1.10
27	NFA021	5	2.05
28	NFA005	5	0.03
29	NFA013	4	0.09
30	NFA011	5	0.27
31	NFA017	4	0.15
32	NFA001	3	0.68
33	NFA007	5	3.86
34	NFA022	5	5.07
35	NFA024	5	6.19

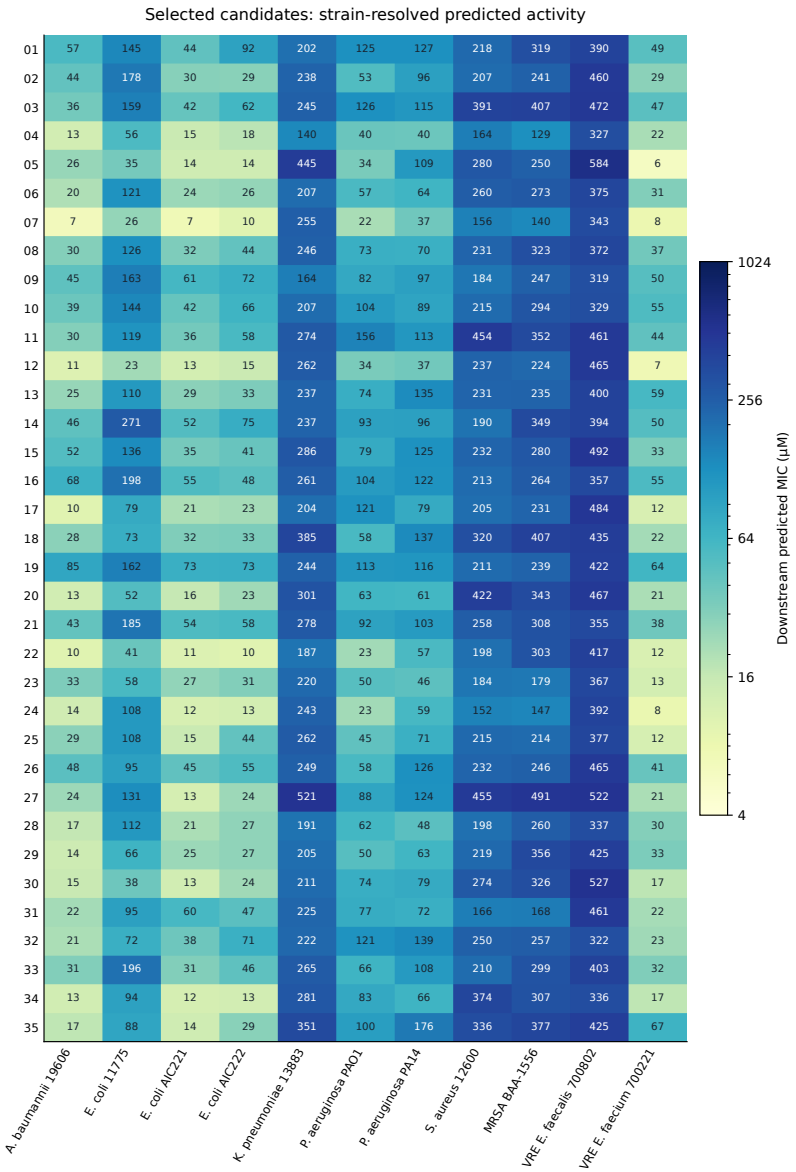


Figure S5: Predicted profiles of the 35 selected candidates. Display IDs are the numeric suffixes of NFAMP identifiers; anchor IDs use the NFA prefix. The left columns identify the fixed anchor, edit distance, and predicted half-life; the heatmap shows downstream strain-resolved predicted MIC values on a logarithmic scale. Candidate and anchor sequences, full-precision objective changes, and screening decisions are supplied in D07. Paired score changes appear in Figure 4; 13 candidates improve all three internal objectives, and 28 improve both safety scores. The half-life estimates are exploratory annotations and were not used for selection.

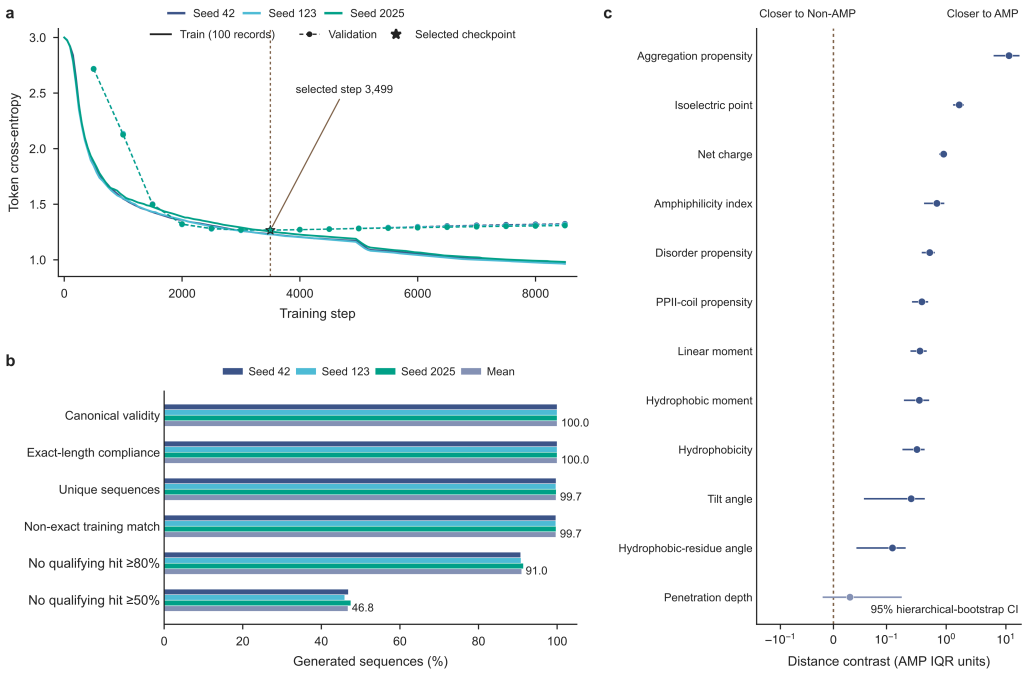


Figure S6: Training and sampling diagnostics. (a) Training token cross-entropy, deterministic validation checks, and selected checkpoints at step 3,499 for each of three seeds; training curves average up to 100 logged observations, recorded approximately every 50 steps. (b) Canonical-residue validity, length compliance, uniqueness, absence of exact training matches, and search-based novelty across 50,000 events per seed. The identity labels denote no qualifying MMseqs2 hit at the stated identity threshold and 80% bidirectional coverage, not an exhaustive global-identity bound. (c) Standardized Wasserstein contrasts for 12 physicochemical properties, with 500-resample hierarchical length-stratified bootstrap intervals. Positive values indicate greater similarity to AMP than Non-AMP references. Panel (c) uses a base-10 symmetric-log axis with linear threshold 0.1 and linear-scale factor 0.8.

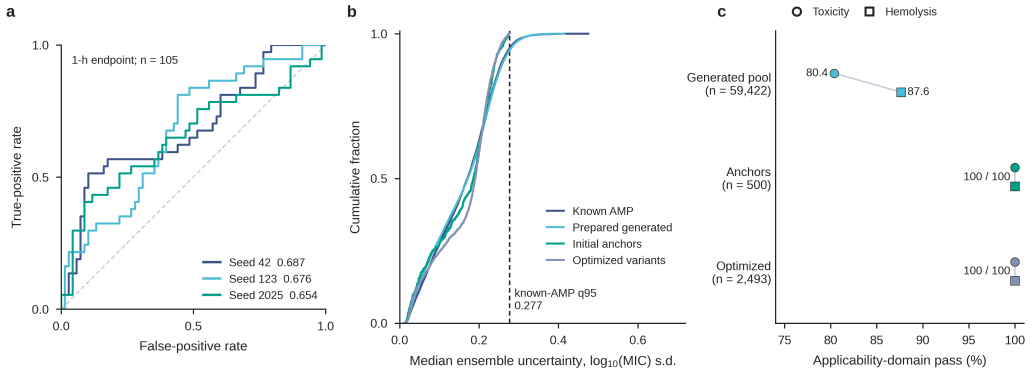


Figure S7: Stability, uncertainty, and applicability. (a) Weighted out-of-fold ROC curves for measured one-hour stability across 105 unique sequences. AUROCs are 0.687, 0.676, and 0.654 for seeds 42, 123, and 2025. (b) Median activity-uncertainty distributions for 16,740 known AMPs, 59,422 prepared generated sequences, 500 anchors, and 2,493 optimized variants. All four distributions use the internal selected-11 APEX scorer. The prepared generated pool has undergone initial preparation; the completely unfiltered sample contains 63,000 events. The dashed line is the known-AMP 95th percentile, 0.277; 56,001 prepared sequences (94.2%) pass it. (c) Safety applicability pass fractions. The reported anchor and final-variant sets pass; this observed property is not guaranteed by the general survivor fallback.

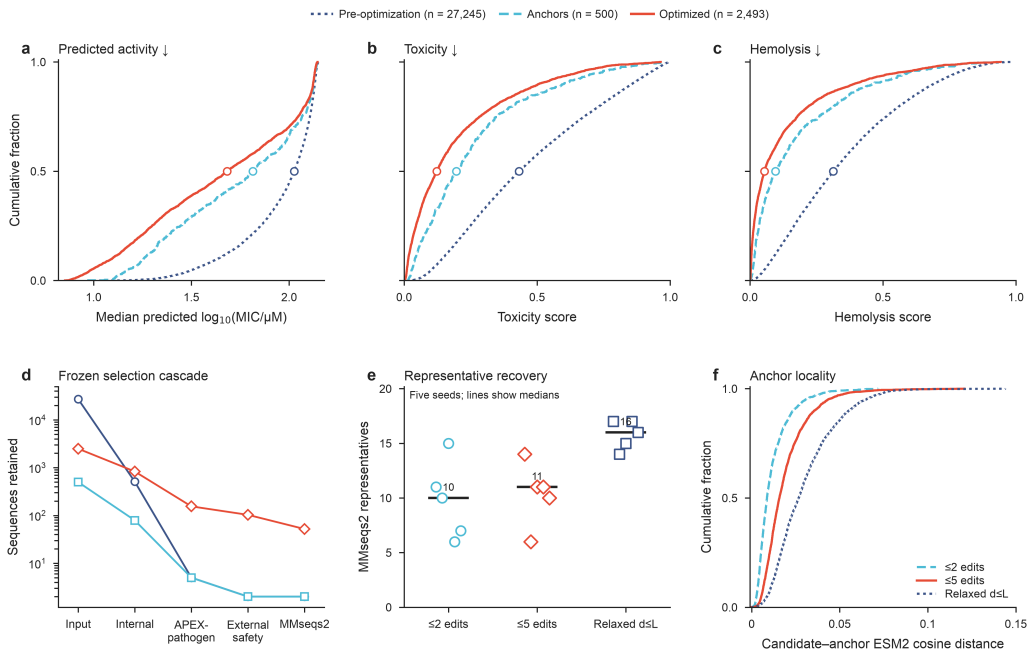


Figure S8: Objective distributions and edit-budget controls. (a–c) Empirical distributions of median predicted $\log_{10}$ MIC, toxicity, and hemolysis for 27,245 feasible pre-optimization sequences, 500 anchors, and 2,493 variants. (d) The same no-half-life-gate cascade retains 2, 2, and 52 representatives, respectively. (e) Per-seed representative counts for the two-edit, five-edit, and relaxed conditions, with medians 10, 11, and 16. (f) Candidate–anchor ESM2 cosine-distance distributions for pooled, sequence-deduplicated final populations ($n = 2,477$, 2,497, and 2,498, rather than 2,500 instances), with medians 0.0087, 0.0158, and 0.0267. These pooled distributions are descriptive sequence-level summaries, distinct from the five seed-level summaries in (e). The relaxed condition uses $B(L) = L$ and retains anchor metadata.

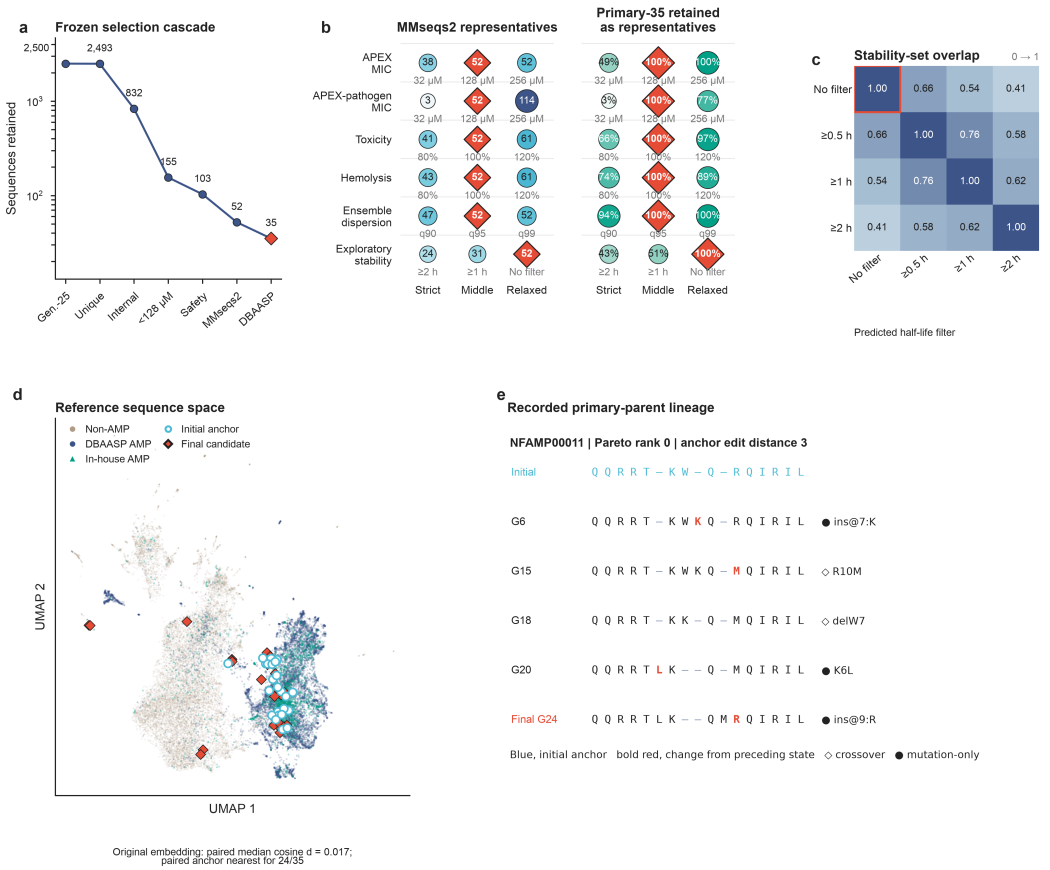


Figure S9: Candidate-selection sensitivity and lineage. (a) Cascade from 2,500 final-population instances to 35 DBAASP-predicted AMP candidates. (b) One-factor sensitivity of representative counts and retention of the primary 35 as representatives; non-half-life rows use no half-life gate. The half-life row compares 2 h, 1 h, and no gate; its primary setting is the rightmost point. Red diamonds mark primary settings. (c) Representative-set Jaccard similarities across exploratory half-life filters; the red frame marks the no-gate primary condition. (d) Reference-fitted UMAP with projected anchors and candidates; quoted distances are calculated in the original normalized 640-dimensional embedding space. (e) Recorded primary-parent path for NFAMP00011, comprising six states with two crossover and three mutation-only transitions. All functional endpoints are predictions. Current-candidate retention means remaining the selected representative, not merely passing the threshold; threshold changes can replace representatives. G24 in (e) denotes creation, followed by survival into generation 25.