CupOFMoCA: Coupled Objective-Guided Discrete Flows for Molecular Conjugate Assembly

Ruoxi Rosie Zhang,[1,*] **Ziang Li,**[2,*] **Jiatao Gu,**[1] **Pranam Chatterjee**[1,3,†]

[1]Department of Computer and Information Science, University of Pennsylvania
[2]College of Arts and Sciences, University of Pennsylvania
[3]Department of Bioengineering, University of Pennsylvania

[*]Equal contribution
[†]Corresponding Author: pranam@upenn.edu
Code: https://huggingface.co/ChatterjeeLab/CupOFMoCA

Abstract

Molecular conjugates, including PROTACs and peptide-drug conjugates (PDCs), derive their function from the joint behavior of multiple coupled components, yet most generative approaches design these components independently and combine them only after generation. Such staged pipelines ignore cross-component dependencies and often produce conjugates that are chemically invalid, fall outside empirical conjugate distributions, or lose function upon assembly. In this work, we introduce **Co**u**p**led **O**bjective-Guided Discrete **F**lows for **Mo**lecular **C**onjugate **A**ssembly (**CupOFMoCA**), a discrete generative method that enforces cross-component chemical constraints during conjugate generation. CupOFMoCA restricts guided updates to chemically feasible conjugate states and reweights local transitions using molecular design objectives. Across three PDC systems, coupled generation achieved 100% conjugation validity among valid peptides and increased valid yield over matched post-hoc filtering, with comparable predicted binding scores across the full valid sets. In three-seed evaluations of three PROTAC systems, CupOFMoCA achieved the highest assembly validity on two systems and generated more assembly-valid molecules per second than the evaluated baselines. We further evaluated the generated PROTACs with post-hoc DegradeMaster scoring and explicit-solvent molecular dynamics, supporting their predicted activity and conformational properties. Together, these results establish CupOFMoCA as an efficient generator of molecular conjugates, supporting therapeutic development via coordinated control of assembly and molecular properties.

1 Introduction

Molecular conjugates play a central role in modern therapeutics and chemical biology by integrating complementary functional components into a single molecular entity Cooper et al. [2021], Hua et al. [2022]. Conjugation strategies improve specificity, modulate pharmacokinetics, and enable mechanisms unavailable to individual components alone [Marei et al., 2022, Eras et al., 2022, Drago et al., 2021]. Representative examples include peptide-drug conjugates (PDCs), which combine sequence-based targeting with small-molecule potency [Rizvi et al., 2024, Heh et al., 2023, Dean et al., 2024], and proteolysis-targeting chimeras (PROTACs), which induce proximity between a target protein and the ubiquitin-proteasome machinery [Vikal et al., 2025, Wang et al., 2024a]. In these systems, function emerges from the coupled behavior of the assembled conjugate rather than from its components in isolation, making conjugate design an inherently joint problem.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Structure-based drug discovery using three-dimensional protein-ligand complexes is widely used Wang et al. [2018], Albanese et al. [2025], but depends on high-quality structural data and is less effective for flexible or disordered systems [Isert et al., 2023, Wang et al., 2018]. Generative modeling has therefore emerged as a complementary approach for molecular design. Autoregressive, masked, and diffusion-based models [Madani et al., 2023, Ferruz et al., 2022, Wang et al., 2024b, Alamdari et al., 2026, Tang et al., 2025a,c, Lee et al., 2025] and discrete flow matching methods [Campbell et al., 2024, Stark et al., 2024, Davis et al., 2024, Tang et al., 2025b, Chen et al., 2025, 2026c] enable efficient generation within individual molecular modalities. However, these approaches typically generate components independently and assemble them post hoc, traversing chemically implausible intermediate states and failing to enforce compatibility at the system level. This limitation is particularly pronounced for conjugates, where validity and function depend on the assembled structure. In PROTACs, for example, linker properties directly control ternary complex formation, and small perturbations can abolish degradation activity despite preserved binding [Békés et al., 2022, Troup et al., 2020, Cao et al., 2022].

To incorporate these dependencies during design, we seek to generate conjugates under joint chemical constraints and molecular objectives. Towards this goal, we introduce **C**o**up**led **O**bjective-Guided Discrete **F**lows for **Mo**lecular **C**onjugate **A**ssembly (**CupOFMoCA**), a discrete generative framework that samples over assembled states under explicit conjugation constraints and target-specific objectives. By operating on discrete flows, CupOFMoCA restricts generation to a chemically feasible conjugate manifold, enabling anticipatory assembly in which intermediate states remain valid and design decisions account for downstream constraints.

Our contributions are threefold:

1. **Coupled discrete flows for molecular conjugate assembly.** We generate conjugates through discrete flows over assembled states and enforce cross-component chemical compatibility during guided sampling.
2. **Objective-guided sampling under hard validity constraints.** We introduce a sampling scheme that incorporates functional objectives while enforcing strict chemical and cross-component validity constraints.
3. **A unified framework across conjugate modalities.** We evaluate CupOFMoCA across three PDC and three PROTAC systems. Coupled PDC generation increased valid yield with comparable predicted binding scores across the full valid sets. For PROTACs, we improved assembly validity through fixed-anchor sampling and linker-specific fine-tuning, then evaluated the generated molecules with post-hoc degradation prediction and molecular dynamics.

2 Related Works

Discrete Flow Matching. Discrete flow matching has emerged as a principled approach for modeling and sampling from complex discrete spaces. One line of work focuses on jump-process models, which learn time-dependent transition rates for token-level stochastic updates and induce continuous-time Markov chains with prescribed marginals [Gat et al., 2024, Campbell et al., 2024]. In contrast, continuous-time simplex methods diffuse discrete data through continuous embeddings over the probability simplex, enabling training via flow-matching or optimal transport objectives defined on relaxed categorical representations [Tang et al., 2025b, Davis et al., 2024, Stark et al., 2024]. More recently, rectified formulations such as Rectified Discrete Flows (ReDi) [Yoo et al., 2025] simplify probability paths and improve sampling efficiency, while AReUReDi [Chen et al., 2026c] extends this paradigm to multi-objective settings with Pareto-guided sampling. Using multi-objective discrete flow matching (MOG-DFM) [Chen et al., 2025] as a base framework, moPPIt establishes motif-specific peptide design [Chen et al., 2026b], while SOAPIA generates target-selective binders with off-target avoidance [Vincoff et al., 2026]. To address viral variation, PROPHET uses MOG-DFM to optimize predicted binding across phylogenetically plausible variants [Liang et al., 2026]. Finally, for editing existing sequences, pCoMole combines preference guidance with hard terminal feasibility [Chen et al., 2026a]. Here, we extend control to cross-component attachment constraints during conjugate sampling.

Constrained Generation for Biomolecule Design. Generative models for biomolecule design have largely focused on optimizing continuous objectives through reward-weighted sampling, classifier

guidance, or guided diffusion [Nisonoff et al., 2025, Gruver et al., 2023]. However, many design requirements are more naturally expressed as feasibility constraints rather than scalar objectives. Examples include representation validity, where generated graphs or SMILES strings must parse into chemically valid molecules [Jin et al., 2018, Tang et al., 2025a, Chen et al., 2026c], and developability thresholds such as thermostability requirements in antibody design [Zeng et al., 2024]. Constraint-aware methods incorporate differentiable constraints into language-model sampling [Kumar et al., 2022] and discrete molecular diffusion [Cardei et al., 2025], while projected diffusion explicitly enforces feasibility [Christopher et al., 2024]. We extend constraint-aware generation to cross-component attachment chemistry in assembled molecular conjugates.

Generative Design of Molecular Conjugates and Linkers. Generative approaches for molecular conjugates have typically focused on individual components. In PROTAC design, SMILES-based methods such as SyntaLinker [Yang et al., 2020], PROTAC-INVENT [Li et al., 2023], and PROTAC-TS [Murakami et al., 2026] treat linker generation as a sequence completion or reinforcement learning problem, while 3D approaches such as DeLinker [Imrie et al., 2020], DiffLinker [Igashov et al., 2024], DiffPROTACs [Li et al., 2024], and LinkerNet [Guan et al., 2023] generate linkers conditioned on fragment geometry. These methods support fragment-conditioned linker design, including ternary-complex docking in PROTAC-INVENT and permeability optimization in PROTAC-TS. CupOFMoCA combines fixed-anchor preservation with explicit assembly checks and objective-guided updates. Similarly, peptide-drug conjugate design pipelines generate peptides independently and attach payloads post hoc, leading to frequent failures at the conjugation stage.

CupOFMoCA integrates discrete generative transport, molecular objectives, and hard cross-component validity checks within a common sampling procedure for PDCs and PROTACs.

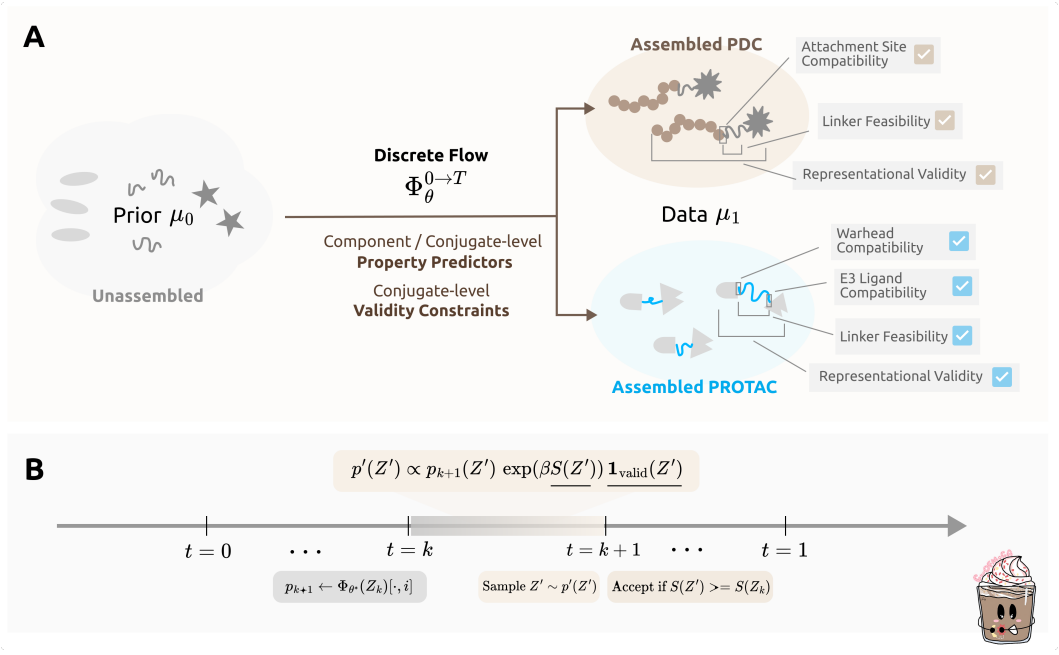

Figure 1: **Overview of CupOFMoCA for molecular conjugate generation. (A)** A pre-trained discrete flow model is guided from unassembled components to valid assembled conjugates, including peptide-drug conjugates (PDCs) and PROTACs, under objective guidance and hard validity constraints (attachment compatibility, linker feasibility, and representational validity). **(B)** Objective-guided sampling with hard validity constraints. We reweight candidate probabilities with objective scores and exclude invalid updates with a validity mask. The illustrated PDC acceptance rule retains updates that preserve or increase the score. For PROTACs, we accept valid tilted proposals and return the highest-scoring valid state encountered.

3 CupOFMoCA

CupOFMoCA generates discrete molecular conjugates by updating free components under constraints imposed by the assembled state. The framework is representation-agnostic, as each component of a conjugate may be expressed as a SMILES token sequence or as a labeled molecular graph, and different components within a single conjugate may use different representations. We instantiate CupOFMoCA in two settings: peptide-drug conjugates (PDCs), where peptides are represented as peptide SMILES strings and the model assembles peptide, linker, and drug components; and the PROTAC-specific instantiation, where states are molecular graphs and the model edits a linker subgraph between fixed warhead and E3-ligand subgraphs (Figure 1A).

3.1 Conjugate State Space

Let $\mathcal{X}^{(c)}$ be the state space of component c, either a SMILES token space $\mathcal{V}^L$ or a graph space $\mathcal{G}_N = \mathcal{A}^N \times \mathcal{B}^{N\times N}$ with empty tokens enabling insertion/deletion. A conjugate is

$$z = \left(x^{(1)}, \ldots, x^{(M)}\right) \in \mathcal{Z} := \textstyle\prod_c \mathcal{X}^{(c)}, \tag{1}$$

with components partitioned into *fixed* ($\mathcal{C}_{\text{fix}}$, immutable inputs) and *free* ($\mathcal{C}_{\text{free}}$, edited by the model). PDCs use $z = (x^{\text{pep}}, x^{\text{link}}, x^{\text{drug}})$ with $\mathcal{C}_{\text{fix}} = \{\text{link}, \text{drug}\}$; PROTACs use $z = (x^{\text{wh}}, x^{\text{link}}, x^{\text{e3}})$ with $\mathcal{C}_{\text{fix}} = \{\text{wh}, \text{e3}\}$. Cross-component bonds at fixed-free junctions are assigned to the adjacent free component.

We assume an unknown data distribution μ_1 over $\mathcal{Z}$ (PDCdb decompositions for PDCs; PROTAC-DB decompositions for PROTACs [Weng et al., 2021]), and a prior μ_0 in which fixed components are set to the target system and free components are drawn from the uninformative DFM prior (uniform random tokens for SMILES, empty subgraph for graphs). The flow learns transport $Z_0 \sim \mu_0 \Rightarrow Z_T \sim \mu_1$ that preserves fixed components throughout.

3.2 Fixed Anchors and Edit Neighborhoods

For each fixed-free junction we specify an *anchor atom* $a^{(c)}$ in $c \in \mathcal{C}_{\text{fix}}$ at which conjugating chemistry occurs; the atom type at $a^{(c)}$ is immutable. In PROTACs on EGFR/VHL, a^{wh} is the Gefitinib-derived -COOH and a^{e3} the VH032 -NH_2, with assembly via amide closure at both anchors. For PDCs, the anchor lives on the fixed linker-drug template (e.g., the Aoa carboxyl in Aoa-Dau, the Cys carboxyl in SSCpt, or the terminal thiol in DSPA-SN38), and the model must generate a peptide exposing a compatible reactive residue (Lys-ε amine or Cys thiol) at the junction. The edit neighborhood

$$\begin{aligned}\mathcal{N}(z) := \big\{ z' \in \mathcal{Z} \,:\, & d_{\text{free}}(z, z') = 1, \\ & x'^{(c)} = x^{(c)} \;\forall\, c \in \mathcal{C}_{\text{fix}} \big\},\end{aligned} \tag{2}$$

admits a single token (SMILES) or atom/bond (graph) substitution on a free component, including bonds incident to anchors (which are themselves immutable).

3.3 Discrete Flows and Transport

We use a *discrete flow* to transport samples from the prior to the data distribution through a sequence of deterministic maps. Let $0 = t_0 < t_1 < \cdots < t_K = T$ denote a discrete time grid. We parameterize a family of transport maps

$$\Phi_\theta^{t_k \to t_{k+1}} : \mathcal{Z} \to \mathcal{Z}, \qquad k = 0, \ldots, K-1, \tag{3}$$

where each map advances a state from time t_k to t_{k+1} while preserving the fixed components and anchor atoms. Starting from $z_0 \sim \mu_0$, the discrete flow evolves by composition,

$$z_{k+1} \;=\; \Phi_\theta^{t_k \to t_{k+1}}(z_k), \tag{4}$$

so that the state at time t_k is given by $z_k = \Phi_\theta^{t_{k-1}\to t_k} \circ \cdots \circ \Phi_\theta^{t_0 \to t_1}(z_0)$. The induced distribution at t_k is the pushforward of the prior,

$$\mu_{t_k}^\theta \;:=\; \left(\Phi_\theta^{t_{k-1}\to t_k} \circ \cdots \circ \Phi_\theta^{t_0 \to t_1}\right)_{\#} \mu_0, \tag{5}$$

and the terminal distribution $\mu_T^\theta = \mu_{t_K}^\theta$ defines the model distribution over assembled conjugates. Each map $\Phi_\theta^{t_k \to t_{k+1}}$ approximates the finite-time transport induced by an underlying continuous probability flow on $\mathcal{Z}$, enabling few-step generation by composing learned transports.

3.4 Hard Constraints and Soft Objectives

We encode validity through a binary indicator $\mathbf{1}_{\text{valid}}(z) \in \{0, 1\}$ that gates sampling: candidates with $\mathbf{1}_{\text{valid}}(z) = 0$ are never accepted. The indicator factorizes into three task-agnostic buckets, each instantiated per conjugation system (Appendix F):

Attachment site compatibility. Conjugation-specific SMARTS checks confirm that the prescribed bond chemistry is realized at every fixed-free junction and that valency at each junction atom is consistent with the reaction (e.g., secondary/tertiary amide N for Lys-ε amide coupling in Aoa-Dau and SSCpt; `SX2`-`SX2` disulfide with no residual free thiol in DSPA-SN38; the analogous warhead/E3 amide patterns in CupOFMoCA). Failure indicators (e.g., unreacted carboxylic acid or free $-$SH) trigger rejection.

Linker feasibility. Conjugation-agnostic structural filters reject infeasible linker regions: strained or oversized ring closures (sp^2-bearing 3-membered rings, 4-membered rings with > 2 sp^2 centers, > 12-atom rings) and excessively long aliphatic runs (≥ 16 consecutive CH_2). For CupOFMoCA these subsume the curated liability filter (hydrazides, azo, N-oxides, N-O singles, carbamates).

Representational validity. Each assembled conjugate must (i) pass standard chemical sanitization, (ii) survive a SMILES parse-serialize round-trip without alteration, (iii) be free of valence and aromaticity violations, and (iv) admit a valid 2D coordinate embedding.

We guide valid proposals with the aggregate score $S(z) = \sum_j w_j S_j(z)$, where S_j is objective j and w_j is its weight. For PDCs, S is the predicted peptide-target binding score used for guidance and evaluation. For PROTACs, we combine synthetic accessibility, chemotype closeness, and connectivity using weights $(4, 2, 4)$ selected through ablation (Appendix J.5). We also apply a separate PROTAC-likeness classifier as a binary validity check.

3.5 Endpoint Couplings

We use a coupling $\gamma \in \Pi(\mu_0, \mu_1)$ that pairs each prior sample with a data conjugate sharing the same fixed components and anchors. Intermediate states z_t are constructed by bounded admissible edits on free components from z_0 toward z_1, yielding training triples (z_0, z_t, z_1).

3.6 Generative Sampling

Following Chen et al. [2026c], generation proceeds in two phases. An unguided phase first denoises samples from the pretrained flow (Model in Appendix E.1, H.1), Φ_{θ^*} until a valid starting state is reached. During guided sampling, we exponentially reweight each per-site distribution with objective scores and exclude invalid candidates with $\mathbf{1}_{\text{valid}}$ (Figure 1B). We accept PDC updates that preserve or increase S. For PROTACs, we accept valid tilted proposals and retain the highest-scoring valid state, following the assembly-validity comparison in Table J3. Implementation details are given in Appendix C for PDC and Appendix J for PROTAC.

4 Peptide-Drug Conjugate Design with CupOFMoCA

For evaluation of CupOFMoCA on PDC design, we focus on three key questions: (1) whether coupled generation guarantees conjugate validity while avoiding functional penalties that could be associated with post-hoc staged conjugation; (2) whether discrete flows with exponential guidance enable high-quality generation; and (3) whether the model exhibits competitive design against therapeutic baselines under *in silico* structural validation.

4.1 Conjugation and Objective Guidance

PDC design conventionally proceeds in a *staged* pipeline: a peptide is first optimized for the target in isolation, then conjugated post hoc to a fixed linker-payload. We hypothesize that this decoupling drives sampling trajectories through chemically infeasible intermediates and overlooks cross-component constraints, including attachment-site compatibility, linker feasibility, and representational validity, that can only be enforced when peptide, linker, and drug evolve jointly. To probe this hypothesis, we benchmarked CupOFMoCA against a broad set of baselines on three therapeutically

relevant targets, each paired with a clinically motivated linker-payload drawn from PDCdb [Sun et al., 2025]: **ERBB2 + SSCpt** (Disulfide-Camptothecin), a validated oncology target in breast and gastric cancers [Harari and Yarden, 2000, Armstrong et al., 2025]; **GNRHR + Aoa-Dau** (Aminooxyacetic acid-Daunorubicin), an endocrine receptor implicated in reproductive disorders and hormone-driven cancers [Maggi et al., 2016]; and **NRP1 + DSPA-SN-38** (3-Disulfanylpropanoic Acid-SN-38), a tumor-vasculature co-receptor [Niland and Eble, 2019] that recognizes C-terminal C-end-rule motifs [Teesalu et al., 2009]. For each target, we enforced conjugation chemistry with the corresponding builder under the hard validity constraints in Appendix F, and we scored peptide-target binding with the affinity predictor used for guidance and report its output in pKd units.

We compared three classes of methods. (i) *Standalone peptide generators*: moPPIt [Chen et al., 2026b] and BoltzGen [Stark et al., 2026] (AA tokenized) and PepTune [Tang et al., 2025a] (SMILES tokenized), each producing target-binding peptides that are conjugated post hoc. (ii) *AReUReDi* [Chen et al., 2026c] in both staged (peptide-only) and guided (conjugation-aware acceptance) modes. (iii) *CupOFMoCA* in the same staged and guided modes. We evaluated AReUReDi and CupOFMoCA under both AA/UAA tokenization with 10-residue peptides (Table 1) and SMILES tokenization with 80- and 100-token peptides (Tables 2, A4).

For AA/UAA generation, we used the pretrained PepReDi peptide prior. PDCdb contains 393 unique canonical peptides within its pretraining range of 6-49 residues, and their zero-shot negative log likelihood was comparable to that of held-out pretraining sequences (Table A2). We therefore retained the broad peptide prior for PDC generation (Appendix A.2).

Table 1: Top-20 *de novo* PDCs across three receptor targets generated by AA-based methods (peptide length: 10 AAs). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker (see Section F). Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). Values are mean (std). **Bold**: best per column; underline: second-best.

	ERBB2		**GNRHR**		**NRP1**	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
moPPIt	7.1%	7.59 (0.07)	8.1%	6.78 (0.46)	65.0%	7.94 (0.06)
BoltzGen						
Exp	51.0%	7.81 (0.05)	16.0%	6.63 (1.05)	-	-
AF	21.2%	7.82 (0.11)	3.0%	7.44 (0.45)	13.7%	7.05 (1.81)
AReUReDi						
Staged (AA)	17.0%	7.65 (0.10)	17.6%	6.58 (0.42)	19.0%	8.23 (0.38)
Guided (AA)	**100.0%**	7.74 (0.09)	**100.0%**	7.15 (0.60)	98.0%	8.81 (0.35)
Staged (UAA)	3.1%	7.68 (0.08)	5.1%	7.68 (0.33)	37.4%	8.07 (0.37)
Guided (UAA)	**100.0%**	**7.83 (0.05)**	**100.0%**	7.93 (0.13)	**100.0%**	8.34 (0.34)
CupOFMoCA						
Staged (AA)	15.0%	7.56 (0.19)	15.0%	6.30 (0.35)	25.0%	8.25 (0.31)
Guided (AA)	**100.0%**	7.75 (0.12)	**100.0%**	7.07 (0.60)	**100.0%**	**8.93 (0.39)**
Staged (UAA)	5.1%	7.77 (0.05)	4.0%	6.98 (0.84)	19.4%	8.02 (0.40)
Guided (UAA)	**100.0%**	7.82 (0.03)	**100.0%**	**8.12 (0.20)**	**100.0%**	8.20 (0.33)

Table 2: Top-20 *de novo* PDCs across three receptor targets generated by SMILES-based methods (peptide length: 80 SMILES tokens). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker.

	ERBB2		**GNRHR**		**NRP1**	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
PepTune	20.8%	6.75 (0.27)	5.6%	4.55	13.7%	7.05 (1.81)
AReUReDi						
Staged	30.1%	7.73 (0.11)	26.8%	7.69 (0.75)	23.0%	8.62 (0.50)
Guided	**100.0%**	**8.02 (0.19)**	**100.0%**	**8.56 (0.35)**	**100.0%**	9.16 (0.36)
CupOFMoCA						
Staged	22.7%	7.73 (0.14)	18.3%	8.09 (1.17)	23.7%	8.76 (0.31)
Guided	**100.0%**	7.97 (0.09)	**100.0%**	8.39 (0.36)	**100.0%**	**9.36 (0.46)**

Coupled generation increased conjugation-valid yield. Under identical endpoint checks, staged generation achieved 3.0-65.0% conjugation validity for AA/UAA methods and 5.6-30.1% for SMILES-80 methods. Guided AReUReDi and CupOFMoCA reached 98-100% across these settings, increasing the fraction of valid peptides that formed valid conjugates (Tables 1 and 2; Figure A1). In the SMILES-80 runs, incompatible attachment chemistry accounted for most staged-generation assembly failures (Figure A2). To assess computational efficiency, we also compared valid-conjugate yield under matched compute. CupOFMoCA-Guided achieved the highest yield per GPU-hour in 8 of 12 target-representation settings (Table A3).

Coupled generation preserved predicted binding. To assess whether the increased valid yield preserved the binding objective, we evaluated all valid conjugates. Across the six SMILES settings, CupOFMoCA increased conjugation validity by 2.7-5.5-fold over its staged counterpart, with full-set mean binding scores within 0.54 pKd units of the corresponding staged means (Tables A6 and A7). The top-20 comparisons and full AA/UAA results further supported retention of predicted binding under conjugation constraints (Tables 1, 2, and A5). To distinguish the contributions of gating and guidance, we compared each component separately. Validity gating raised conjugation validity to 100%, and affinity guidance increased mean predicted binding by approximately 0.3-1.1 pKd units within the feasible region (Table A1).

Tokenization regimes. To examine how representation affected design, we tested AA, UAA, and SMILES tokenizations. AA covers the 20 canonical residues, UAA adds non-canonical monomers, and SMILES operates at the atom level. Predicted binding increased with expressivity for both AReUReDi and CupOFMoCA, with the highest pKd in the SMILES setting on every target. CupOFMoCA-Guided reached 100% conjugation validity across these regimes and matched AReUReDi-Guided overall, with higher predicted binding on NRP1 in the AA and SMILES settings (Tables 1, 2). We observed the same pattern with 100-token SMILES (Table A4).

Molecular properties and linker diversity. Generated PDCs spanned a physicochemical regime similar to PDCdb pharmaceutical references on molecular weight, LogP, TPSA, and heavy-atom count, with target-specific distributions reported in Appendix A (Figure A3; per-target breakdowns in Tables A10-A18). To assess flexibility beyond fixed linkers, we sampled linker types from PDCdb frequencies. The resulting PDCs matched the binding and validity of the fixed-linker setting (Table A19), with the expected linker-type distribution (Table A20). Across all settings, we obtained high uniqueness and diversity with low training-set similarity (Table A21).

5 PROTAC Design with CupOFMoCA

5.1 Constraints and Objective Guidance

PROTAC assembly imposes hard validity constraints coupling the linker with the fixed warhead and E3 ligand, encoded via a binary classifier $\mathbf{1}_{\text{valid}}(z) \in \{0, 1\}$ that returns 1 iff z satisfies all of:

(a) **Representational validity.** The molecule passes RDKit sanitization and the SMILES round-trip $\text{MolFromSmiles}(\text{MolToSmiles}(z))$ recovers a valid molecule; states with malformed valences or aromaticity are rejected.

(b) **Liability filter.** No SMARTS match against hydrazides (`[NX3][NX3]`), azo groups (`[#7]=[#7]`), N-oxides (`[N+][O-]`), N-O single bonds (`[#7]-[OX2]`), or carbamates (`[NX3]C(=[OX1])[OX2H1]`), which decompose, exhibit toxicity, or fail amide-coupled assembly.

(c) **Reactive termini count.** At least two carbon-bonded carboxylic acids (`[#6][CX3](=[OX1])[OX2H1]`) or aldehydes (`[#6][CX3H1](=[OX1])`); fewer renders warhead/E3 amide-coupling impossible.

Candidates with $\mathbf{1}_{\text{valid}}(z) = 0$ receive zero probability under the sampling distribution.

For more details on the validity constraints, see Appendix I. Beyond the hard gate, we use a continuous aggregate objective

$$S(z) \;=\; \sum_{j=1}^{3} w_j \, S_j(z) \;\in\; \mathbb{R}, \tag{6}$$

combining three soft scorers with default weights $(w_1, w_2, w_3) = (4.0,\ 2.0,\ 4.0)$, selected through the weight sweeps in Appendix J.5; all PROTAC results in Section 5.2 use them:

S_1 (**Synthetic accessibility**): a linear ramp on RDKit's synthetic accessibility (SA) score, 1.0 at $\mathrm{SA} = 1$ and 0.0 at $\mathrm{SA} = 6$, with $\mathrm{SA} > 6$ mapped to zero.

S_2 (**Chemotype closeness**): a trapezoidal windowed score from PROTAC-DB statistics averaging ether count, $\log P$, linker MW, and rotatable-bond count, each mapped into $[0, 1]$ via a window centered on the empirical PROTAC-DB range (assembled mass $[771, 1108]$ Da minus fixed warhead/E3 offsets).

S_3 (**Connectivity**): returns 1.0 for a single connected component, otherwise $\frac{|\text{largest fragment}|}{|\text{total atoms}| \cdot |\text{fragments}|}$, halved if no path exists through x^{link} between a^{wh} and a^{e3}. We evaluate S_3 on partially assembled states with $\mathbf{1}_{\text{valid}}(z) = 0$. The elevated w_3 gives connected, anchor-bridging linkers greater weight early in transport.

With this aggregate score, we favor connected linkers with lower SA scores and PROTAC-DB-like properties. We evaluated the objective weights and the joint contribution of hard constraints and soft guidance in Appendix J.5 and Table J8.

5.2 PROTAC Linker Generation

We benchmarked CupOFMoCA against LinkerNet [Guan et al., 2023], DiffPROTACs [Li et al., 2024], and DiffLinker [Igashov et al., 2024] on three PROTAC systems (EGFR-VHL, BRD9-VHL, CDK6-CRBN). For CupOFMoCA, we generated 100 candidates per system with each of three random seeds; baseline results use one batch of 100 candidates per system. We evaluated every candidate with the same attachment, valence, liability, and assembly checks, without method-specific repair (Appendix I). We report standard and assembly validity with post-hoc predicted degradation (Table 3), and diversity and training-set similarity (Table 4).

Table 3: Assembly validity and predicted degradation across PROTAC systems. Standard validity is the fraction of RDKit-parseable molecules; assembly validity additionally requires canonical attachment and an analyzer pass. With the selected objective weights (Appendix J.5), we generated 100 candidates for each of three seeds (42, 123, 456), obtaining 255, 247, and 249 assembly-valid molecules for EGFR-VHL, BRD9-VHL, and CDK6-CRBN, respectively. Validity is mean ± std across seeds; each baseline used one batch of 100. DegradeMaster scores are mean (std) over assembly-valid molecules, and threshold columns report the fraction scoring ≥ 0.5. **Bold**: best per column; underline: second-best. Dashes denote no valid assemblies.

	EGFR-VHL		**BRD9-VHL**		**CDK6-CRBN**	
Method	Std. Val.	Asm. Val.	Std. Val.	Asm. Val.	Std. Val.	Asm. Val.
LinkerNet	**98.0**	46.0	**100.0**	16.0	**100.0**	**96.0**
DiffPROTACs	**98.0**	80.0	**100.0**	0.0	86.0	74.0
DiffLinker	86.0	0.0	74.0	0.0	92.0	0.0
Ours	95.3 ± 2.1	$\mathbf{85.0 \pm 3.6}$	94.3 ± 2.1	$\mathbf{82.3 \pm 4.0}$	94.7 ± 1.5	83.0 ± 4.2
	DegradeMaster predicted degradation probability					
	EGFR-VHL		**BRD9-VHL**		**CDK6-CRBN**	
Method	mean	≥ 0.5	mean	≥ 0.5	mean	≥ 0.5
LinkerNet	0.278 (0.051)	0.0%	0.172 (0.053)	0.0%	**0.925 (0.055)**	**100.0%**
DiffPROTACs	0.306 (0.236)	20.0%	-	-	0.882 (0.185)	96.7%
Ours	**0.509 (0.275)**	**52.3%**	**0.260 (0.235)**	**18.9%**	0.770 (0.282)	80.4%

Standard and assembly validity. Table 3 distinguishes *standard validity* (RDKit-parseable SMILES) from *assembly validity* (canonical attachment and analyzer pass). Under the common

endpoint analyzer, DiffLinker produced parseable SMILES at 74-92% but no valid assemblies. LinkerNet reached 46.0%, 16.0%, and 96.0% assembly validity on EGFR-VHL, BRD9-VHL, and CDK6-CRBN, while DiffPROTACs reached 80.0%, 0.0%, and 74.0%. With fixed anchors (Section 3.2) and validity-gated sampling, CupOFMoCA reached $85.0 \pm 3.6\%$, $82.3 \pm 4.0\%$, and $83.0 \pm 4.2\%$, ranking first on EGFR-VHL and BRD9-VHL and second on CDK6-CRBN. Under matched hardware, CupOFMoCA generated 0.247 assembly-valid PROTACs per second, compared with 0.181 for DiffPROTACs, 0.072 for LinkerNet, and 0.000 for DiffLinker (Appendix J.6).

Predicted degradation and physical evaluation. We evaluated assembly-valid PROTACs with DegradeMaster [Liu et al., 2025], which we applied after generation for in-domain activity prediction. DegradeMaster and the fine-tuned generator both draw on PROTAC databases. CupOFMoCA reached mean predicted probabilities of 0.509, 0.260, and 0.770 on EGFR-VHL, BRD9-VHL, and CDK6-CRBN; 52.3%, 18.9%, and 80.4% of its assembly-valid generations scored ≥ 0.5 (Table 3). CupOFMoCA had the highest predicted probability on EGFR-VHL and BRD9-VHL, whereas LinkerNet performed best on CDK6-CRBN.

To complement the degradation predictions with a physical assessment, we evaluated four generated EGFR-VHL candidates by explicit-solvent molecular dynamics using MS39 as a cell-active positive control and a PEG4-linked congener as a negative control. The generated candidates occupied a conformational regime comparable to MS39, and C1 and C2 had lower 3D polar surface area and lower ternary-compatible conformational penalties. We detail the simulation protocol and compound definitions in Appendix B.3 and Table B2. The comparison with both controls supported favorable linker-dependent conformational properties.

Table 4: Diversity and training-set similarity across PROTAC systems. Div: mean pairwise Tanimoto within the generated batch (lower = more diverse). Sim: mean max-Tanimoto to PROTAC-DB (lower = less similar to training data). "Sim >0.8" is the fraction exceeding the similarity threshold. Our rows pool three seeds (255 / 247 / 249 assembly-valid molecules); baselines are single batches of $N = 100$. **Bold**: best per column; underline: second-best. Dashes indicate methods that produce 0% valid assemblies.

	EGFR-VHL			BRD9-VHL			CDK6-CRBN		
Method	Div	Sim	>0.8	Div	Sim	>0.8	Div	Sim	>0.8
LinkerNet	0.895	0.816	78.3%	0.829	0.825	75.0%	0.833	0.776	45.8%
DiffPROTACs	**0.763**	0.779	35.0%	-	-	-	0.735	0.694	2.7%
Ours	0.772	**0.630**	**0.0%**	**0.746**	**0.777**	**26.0%**	**0.716**	**0.545**	**0.0%**

Diversity and training-set similarity. Table 4 reports mean pairwise Tanimoto similarity within each generated batch and maximum Tanimoto similarity to PROTAC-DB. CupOFMoCA had the lowest mean maximum training-set similarity on all three systems. None of its EGFR-VHL or CDK6-CRBN generations exceeded the 0.80 similarity threshold, while 26.0% of its BRD9-VHL generations did. PROTAC-DB contains 37 reference compounds for BRD9-VHL, 200 for EGFR-VHL, and 117 for CDK6-CRBN. CupOFMoCA had the highest diversity on BRD9-VHL and CDK6-CRBN and the second-highest diversity on EGFR-VHL. On CDK6-CRBN, LinkerNet combined a mean DegradeMaster score of 0.925 with 45.8% of generations above the similarity threshold; CupOFMoCA reached 0.770, with no generations above the threshold. The activity predictions and similarity analysis therefore characterized complementary properties of the generated candidates.

Takeaways. Across the three evaluated systems, fixed anchors (Section 3.2), hard validity constraints, and soft guidance (Section 5.1) supported high assembly-valid yield under a common endpoint analyzer, with the strongest performance from their combination (Table J8). PROTAC-specific fine-tuning of the graph flow on PROTAC-DB linkers raised assembly validity by 4.7-6.3 percentage points and lowered mean SA from approximately 3.6 to 2.8 (Table H1), and CupOFMoCA achieved the highest assembly validity on two systems and the second-highest value on one. The generated PROTACs contained diverse linkers between the fixed warhead and E3 ligand (Figure B1), with target-specific molecular properties summarized in Table B1.

6 Discussion

In this work, we present **CupOFMoCA**, a method for objective-guided discrete generation under coupled chemical constraints. During guided sampling, we enforce chemical compatibility between components so that intermediate states correspond to valid partial conjugates. Applied to peptide-drug conjugates (PDCs), this formulation produced fully assembled molecules, increased conjugation-valid yield relative to staged baselines, and maintained predicted binding scores near the corresponding staged levels. Applied to PROTACs, the same fixed-anchor and hard-validity machinery reduced the gap between standard and assembly validity, achieving the highest assembly validity on two of the three evaluated systems and the second-highest value on the third. Beyond the sampling objectives, post-hoc in-domain DegradeMaster scoring and explicit-solvent molecular dynamics provided complementary computational evaluations of completed candidates. CupOFMoCA also produced linkers with the lowest mean maximum similarity to training compounds on all three systems.

While we used molecular dynamics to characterize linker conformations and ternary-geometry compatibility, we have yet to establish the biological activity of our generated candidates experimentally. To connect these computational properties to cellular function, we are currently performing full ternary-complex simulations and wet lab experiments to measure stability, target engagement, and degradation. Beyond these evaluations, we will next extend CupOFMoCA to antibody-drug conjugates and multicomponent degraders, incorporate pharmacokinetic and linker-property objectives, and enable adaptive selection of linker-payload combinations. With these extensions, we envision that CupOFMoCA will enable the next-generation of molecular conjugates, optimized for coordinated control over component compatibility and therapeutic function.

Author Contributions

R.R.Z. formulated and trained CupOFMoCA and conducted all PDC experiments. Z.L. formulated and conducted all PROTAC experiments. J.G. supervised the research and reviewed the manuscript. P.C. conceived, formulated, designed, directed, and supervised the study.

Funding Statement.

The work was funded by a grant to the lab of P.C. from the Futures Fund at the University of Pennsylvania.

Competing Interests.

P.C. is a co-founder of Gameto, Inc., UbiquiTx, Inc., AtomBioworks, Inc., and Kosha Sciences, Inc. and advises companies involved in therapeutics development. P.C.'s interests are reviewed and managed by the University of Pennsylvania in accordance with their conflict-of-interest policies. The remaining authors have no conflicts of interest to declare.

References

Osama Abdin, Satra Nim, Han Wen, and Philip M Kim. PepNN: a deep attention model for the identification of peptide binding sites. *Communications Biology*, 5(1):503, 2022. doi: 10.1038/s42003-022-03445-2.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330701. URL `https://doi.org/10.1145/3292500.3330701`.

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M. Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: Sequence is all you need. *eLife*, 2026. doi: 10.7554/eLife.112029.1. URL

https://elifesciences.org/reviewed-preprints/112029. Reviewed preprint, version 1.

Katherine I Albanese, Sophie Barbe, Shunsuke Tagami, Derek N Woolfson, and Thomas Schiex. Computational protein design. *Nature Reviews Methods Primers*, 5(1):13, 2025. doi: 10.1038/s43586-025-00383-1.

Amy Armstrong, Fleur Coburn, Yanyamba Nsereko, and Othman Al Musaimi. Peptide-Drug Conjugates: A New Hope for Cancer. *Journal of Peptide Science*, 31(8):e70040, 2025. doi: 10.1002/psc.70040.

Miklós Békés, David R. Langley, and Craig M. Crews. PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery*, 21(3):181–200, 2022. doi: 10.1038/s41573-021-00371-6.

Anton Bushuiev, Roman Bushuiev, Petr Kouba, Anatolii Filkin, Markéta Gabrielová, Michal Gabriel, Jiri Sedlar, Tomas Pluskal, Jiri Damborsky, Stanislav Mazurenko, and Josef Sivic. Learning to design protein-protein interactions with enhanced generalization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5b4263be85820683d78675cc18d2efc7-Abstract-Conference.html.

Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative Flows on Discrete State-Spaces: Enabling Multimodal Flows with Applications to Protein Co-Design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5453–5512. PMLR, 2024. URL https://proceedings.mlr.press/v235/campbell24a.html.

Chaoguo Cao, Ming He, Liguo Wang, Yuna He, and Yu Rao. Chemistries of bifunctional PROTAC degraders. *Chemical Society Reviews*, 51(16):7066–7114, 2022. doi: 10.1039/D2CS00220E.

Michael Cardei, Jacob K. Christopher, Thomas Hartvigsen, Bhavya Kailkhura, and Ferdinando Fioretto. Constrained Molecular Generation with Discrete Diffusion for Drug Discovery. In *NeurIPS 2025 Workshop on AI Virtual Cells and Instruments: A New Era in Drug Discovery and Development*, 2025. URL https://ai4d3.github.io/2025/papers/18_Constrained_Molecular_Gener.pdf.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-Objective-Guided Discrete Flow Matching for Controllable Biological Sequence Design. In *ICML Workshop on Generative AI and Biology (GenBio)*, 2025. URL https://openreview.net/forum?id=8YIMLoHP9J.

Tong Chen, Maximilian Holsman, Lin Zhao, and Pranam Chatterjee. pCoMole: Pareto-Constrained Molecule Editing with Discrete Flows. In *Advances in Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=IUuHJ0x69N.

Tong Chen, Zachary Quinn, Kunal Mishra, Erin C. O'Connor, Sophia E. Silver, Yinuo Zhang, Mary Jo Valencia, Ying Mei, Jacques Behmoaras, Leonardo M. R. Ferreira, and Pranam Chatterjee. moPPIt: De Novo Generation of Motif-Specific and Functionally Active Peptide Binders via Discrete Flow Matching. *bioRxiv*, 2026b. doi: 10.1101/2024.07.31.606098. URL https://www.biorxiv.org/content/10.1101/2024.07.31.606098v2. Preprint, version 2.

Tong Chen, Yinuo Zhang, Sophia Vincoff, and Pranam Chatterjee. AReUReDi: Annealed Rectified Updates for Refining Discrete Flows with Multi-Objective Guidance. In *ICLR Workshop on Real-World Constrained and Preference-Aligned Flow and Diffusion-Based Models*, 2026c. URL https://openreview.net/forum?id=kugFCaacGz.

Meng Cheng, Xufen Yu, Kaylene Lu, Ling Xie, Li Wang, Fanye Meng, Xiaoran Han, Xian Chen, Jing Liu, Yue Xiong, and Jian Jin. Discovery of potent and selective epidermal growth factor receptor (EGFR) bifunctional small-molecule degraders. *Journal of Medicinal Chemistry*, 63(3):1216–1232, 2020. doi: 10.1021/acs.jmedchem.9b01566.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pages 89307–89333, 2024. doi: 10.52202/079017-2834.

URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a28af221f2f70be183afc16797a56b91-Abstract-Conference.html.

Bethany M Cooper, Jessica Iegre, Daniel H O'Donovan, Maria Ölwegård Halvarsson, and David R Spring. Peptides as a platform for targeted therapeutics for cancer: peptide-drug conjugates (PDCs). *Chemical society reviews*, 50(3):1480–1494, 2021. doi: 10.1039/D0CS00556H.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pages 139054–139084, 2024. doi: 10.52202/079017-4413. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/fadec8f2e65f181d777507d1df69b92f-Abstract-Conference.html.

Trevor T Dean, Juliet Jelu-Reyes, A'Lester C Allen, and Terry W Moore. Peptide-drug conjugates: An emerging direction for the next generation of peptide therapeutics. *Journal of Medicinal Chemistry*, 67(3):1641–1661, 2024. doi: 10.1021/acs.jmedchem.3c01835.

Joshua Z Drago, Shanu Modi, and Sarat Chandarlapaty. Unlocking the potential of antibody-drug conjugates for cancer therapy. *Nature Reviews Clinical Oncology*, 18(6):327–344, 2021. doi: 10.1038/s41571-021-00470-8.

Alexis Eras, Danna Castillo, Margarita Suárez, Nelson Santiago Vispo, Fernando Albericio, and Hortensia Rodriguez. Chemical conjugation in drug delivery systems. *Frontiers in Chemistry*, 10: 889083, 2022. doi: 10.3389/fchem.2022.889083.

Aaron L Feller and Claus O Wilke. p2smi: A toolkit enabling SMILES generation and property analysis for noncanonical and cyclized peptides. *Journal of Open Source Software*, 10(116):8319, 2025a. doi: 10.21105/joss.08319.

Aaron L Feller and Claus O Wilke. Peptide-aware chemical language model successfully predicts membrane diffusion of cyclic peptides. *Journal of Chemical Information and Modeling*, 65(2): 571–579, 2025b. doi: 10.1021/acs.jcim.4c01441.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13:4348, 2022. doi: 10.1038/s41467-022-32007-7. URL https://www.nature.com/articles/s41467-022-32007-7.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37, pages 133345–133385, 2024. doi: 10.52202/079017-4239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0d629a734b56a642701bba7bc8bb3ed-Abstract-Conference.html.

Nate Gruver, Samuel Stanton, Nathan Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein Design with Guided Discrete Diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pages 12489–12517, 2023. doi: 10.52202/075280-0547. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/29591f355702c3f4436991335784b503-Abstract-Conference.html.

Jiaqi Guan, Xingang Peng, PeiQi Jiang, Yunan Luo, Jian Peng, and Jianzhu Ma. LinkerNet: Fragment poses and linker co-design with 3D equivariant diffusion. In *Advances in Neural Information Processing Systems*, volume 36, pages 77503–77519, 2023. doi: 10.52202/075280-3388. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4821075019a058700f6e6738eea1365-Abstract-Conference.html.

Daniel Harari and Yosef Yarden. Molecular mechanisms underlying ErbB2/HER2 action in breast cancer. *Oncogene*, 19(53):6102–6114, 2000. doi: 10.1038/sj.onc.1203973.

Ethan Heh, Jesse Allen, Fabiola Ramirez, Daniel Lovasz, Lorena Fernandez, Tanis Hogg, Hannah Riva, Nathan Holland, and Jessica Chacon. Peptide drug conjugates and their role in cancer therapy. *International Journal of Molecular Sciences*, 24(1):829, 2023. doi: 10.3390/ijms24010829.

Liwen Hua, Qiuyue Zhang, Xinyue Zhu, Ruoning Wang, Qidong You, and Lei Wang. Beyond proteolysis-targeting chimeric molecules: designing heterobifunctional molecules based on functional effectors. *Journal of Medicinal Chemistry*, 65(12):8091–8112, 2022. doi: 10.1021/acs.jmedchem.2c00316.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6(4):417–427, 2024. doi: 10.1038/s42256-024-00815-9.

Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3D linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995, 2020. doi: 10.1021/acs.jcim.9b01120.

Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023. doi: 10.1016/j.sbi.2023.102548.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 2018. URL https://proceedings.mlr.press/v80/jin18a.html.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.144. URL https://aclanthology.org/2022.emnlp-main.144/.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, NadineSchneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, tadhurst-cdd, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Rachel Walker, Vincent F. Scalfani, Hussein Faara, Kazuya Ujihara, Daniel Probst, Juuso Lehtivarjo, Guillaume Godin, Axel Pahl, and Jeremy Monat. rdkit/rdkit: 2025_03_1 (Q1 2025) Release, 2025. URL https://zenodo.org/records/15115844. Software, version Release_2025_03_1.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Gopal Paliwal, Weili Nie, and Arash Vahdat. GenMol: A Drug Discovery Generalist with Discrete Diffusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 33205–33226. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25o.html.

Baiqing Li, Ting Ran, and Hongming Chen. 3D-based generative PROTAC linker design with reinforcement learning. *Briefings in Bioinformatics*, 24(5):bbad323, 2023. doi: 10.1093/bib/bbad323.

Fenglei Li, Qiaoyu Hu, Yongqi Zhou, Hao Yang, and Fang Bai. DiffPROTACs is a deep learning-based generator for proteolysis targeting chimeras. *Briefings in Bioinformatics*, 25(5):bbae358, 2024. doi: 10.1093/bib/bbae358.

Kimberly Liang, Navya Nori, Sophia Vincoff, Elizabeth H. Mahood, Yinuo Zhang, and Pranam Chatterjee. PROPHET: Phylogenetically Robust Antiviral Peptide Design Against Heterogeneous Evolutionary Trajectories. In *ICML Workshop on Generative and Agentic AI for Biology*, 2026. URL https://openreview.net/forum?id=zQJ0Ll2IR5.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Jie Liu, Michael J. Roy, Luke Isbel, and Fuyi Li. Accurate PROTAC-targeted degradation prediction with DegradeMaster. *Bioinformatics*, 41(Supplement_1):i342–i351, 2025. doi: 10.1093/bioinformatics/btaf191.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2. URL https://www.nature.com/articles/s41587-022-01618-2.

Roberto Maggi, Anna Maria Cariboni, Marina Montagnani Marelli, Roberta Manuela Moretti, Valentina Andrè, Monica Marzagalli, and Patrizia Limonta. GnRH and GnRH receptors in the pathophysiology of the human female reproductive system. *Human Reproduction Update*, 22(3): 358–381, 2016. doi: 10.1093/humupd/dmv059.

Hany E Marei, Carlo Cenciarelli, and Anwarul Hasan. Potential of antibody-drug conjugates (ADCs) for cancer therapy. *Cancer Cell International*, 22(1):255, 2022. doi: 10.1186/s12935-022-02679-8.

Yuki Murakami, Shoichi Ishida, Nobuo Cho, Hitomi Yuki, Masateru Ohta, Teruki Honma, Yosuke Demizu, and Kei Terayama. Data-driven design of PROTAC linkers to improve PROTAC cell membrane permeability. *JACS Au*, 6(2):1400–1410, 2026. doi: 10.1021/jacsau.6c00033.

Stephan Niland and Johannes A Eble. Neuropilins in the context of tumor vasculature. *International Journal of Molecular Sciences*, 20(3):639, 2019. doi: 10.3390/ijms20030639.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/597254dc45be8c166d3ccf0ba2d56325-Abstract-Conference.html.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707. Preprint.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Vasanthanathan Poongavanam, Yoseph Atilaw, Stephan Siegel, Anja Giese, Lutz Lehmann, Daniel Meibom, Máte Erdelyi, and Jan Kihlberg. Linker-dependent folding rationalizes PROTAC cell permeability. *Journal of Medicinal Chemistry*, 65(19):13029–13040, 2022. doi: 10.1021/acs.jmedchem.2c00877.

Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. DeFoG: Discrete Flow Matching for Graph Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 50269–50326. PMLR, 2025. URL https://proceedings.mlr.press/v267/qin25d.html.

Syed Faheem Askari Rizvi, Linjie Zhang, Haixia Zhang, and Quan Fang. Peptide-drug conjugates: design, chemistry, and drug delivery system as a novel cancer theranostic. *ACS Pharmacology & Translational Science*, 7(2):309–334, 2024. doi: 10.1021/acsptsci.3c00269.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 130136–130184, 2024. doi: 10.52202/079017-4135. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet Flow Matching with Applications to DNA Sequence Design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 46495–46513. PMLR, 2024. URL https://proceedings.mlr.press/v235/stark24b.html.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tally Portnoi, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, Samuel J. Snyder, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Haoyu Fan, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward Universal Binder Design. *bioRxiv*, 2026. doi: 10.1101/2025.11.20.689494. URL https://www.biorxiv.org/content/10.1101/2025.11.20.689494v2. Preprint, version 2.

Xiuna Sun, Hanyang Li, Zhen Chen, Yang Zhang, Zhangle Wei, Hangwei Xu, Yang Liao, Wanghao Jiang, Yichao Ge, Lingyan Zheng, Teng Li, Yuting Wu, Meiyin Luo, Luo Fang, Xiaowu Dong, Mang Xiao, Lianyi Han, Qingzhong Jia, and Feng Zhu. PDCdb: the biological activity and pharmaceutical information of peptide-drug conjugate (PDC). *Nucleic Acids Research*, 53(D1): D1476–D1485, 2025. doi: 10.1093/nar/gkae859.

Sophia Tang, Yinuo Zhang, and Pranam 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 of *Proceedings of Machine Learning Research*, pages 59017–59065. PMLR, 2025a. URL https://proceedings.mlr.press/v267/tang25n.html.

Sophia Tang, Yinuo Zhang, Alexander Tong, and Pranam Chatterjee. Gumbel-Softmax Score and Flow Matching for Discrete Biological Sequence Generation. In *ICLR Workshop on Frontiers in Probabilistic Inference*, 2025b. URL https://openreview.net/forum?id=vx1uo7AAtW.

Sophia Tang, Yuchen Zhu, Molei Tao, and Pranam Chatterjee. TR2-D2: Tree Search Guided Trajectory-Aware Fine-Tuning for Discrete Diffusion. *arXiv preprint arXiv:2509.25171*, 2025c. URL https://arxiv.org/abs/2509.25171.

Tambet Teesalu, Kazuki N Sugahara, Venkata Ramana Kotamraju, and Erkki Ruoslahti. C-end rule peptides mediate neuropilin-1-dependent cell, vascular, and tissue penetration. *Proceedings of the National Academy of Sciences*, 106(38):16157–16162, 2009. doi: 10.1073/pnas.0908201106.

Robert I. Troup, Charlene Fallan, and Matthias G. J. Baud. Current strategies for the design of PROTAC linkers: a critical review. *Exploration of Targeted Anti-Tumor Therapy*, 1(5):273–312, 2020. doi: 10.37349/etat.2020.00018.

Akash Vikal, Rashmi Maurya, Brij Bihari Patel, Rajeev Sharma, Preeti Patel, Umesh K Patil, and Balak Das Kurmi. Protacs in cancer therapy: mechanisms, design, clinical trials, and future directions. *Drug Delivery and Translational Research*, 15(6):1801–1827, 2025. doi: 10.1007/s13346-024-01754-z.

Sophia Vincoff, Howard Liu, Corinne M. Linardic, and Pranam Chatterjee. SOAPIA: Specificity-Guided Generation of Off-Target-Avoiding Protein Interactions with High Target Affinity. In *ICML Workshop on Generative and Agentic AI for Biology*, 2026. URL https://openreview.net/forum?id=qN3sOnGDmK.

Chao Wang, Yujing Zhang, Wujun Chen, Yudong Wu, and Dongming Xing. New-generation advanced PROTACs as potential therapeutic agents in cancer therapy. *Molecular Cancer*, 23(1): 110, 2024a. doi: 10.1186/s12943-024-02024-9.

Xin Wang, Ke Song, Li Li, and Lijiang Chen. Structure-based drug design strategies and challenges. *Current Topics in Medicinal Chemistry*, 18(12):998–1006, 2018. doi: 10.2174/1568026618666180813152921.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion Language Models Are Versatile Protein Learners. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52309–52333. PMLR, 2024b. URL https://proceedings.mlr.press/v235/wang24ct.html.

Gaoqi Weng, Chao Shen, Dongsheng Cao, Junbo Gao, Xiaowu Dong, Qiaojun He, Bo Yang, Dan Li, Jian Wu, and Tingjun Hou. PROTAC-DB: an online database of PROTACs. *Nucleic Acids Research*, 49(D1):D1381–D1387, 2021. doi: 10.1093/nar/gkaa807.

Yuyao Yang, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen. SyntaLinker: automatic fragment linking with deep conditional transformer neural networks. *Chemical Science*, 11(31):8312–8322, 2020. doi: 10.1039/d0sc03126g.

Jaehoon Yoo, Wonjung Kim, and Seunghoon Hong. ReDi: Rectified Discrete Flow. In *Advances in Neural Information Processing Systems*, volume 38, pages 81651–81683, 2025. doi: 10.52202/085713-2727. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7526508f11bbe0a123af62b9dab1fbe1-Abstract-Conference.html.

Yimeng Zeng, Hunter Elliott, Phillip Maffettone, Peyton Greenside, Osbert Bastani, and Jacob R Gardner. Antibody design with constrained bayesian optimization. In *ICLR Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024. URL https://openreview.net/forum?id=K5Sr6WSA4B.

Chengxin Zhang, Xi Zhang, Lydia Freddolino, and Yang Zhang. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024. doi: 10.1093/nar/gkad630.

Appendix

A PDC Additional Results

A.1 Validity-Gate and Affinity-Guidance Ablation

We isolated the roles of the learned flow, conjugation-validity gate, and affinity guidance by independently enabling the latter two components. The *unconditional* condition uses neither; *staged* uses affinity guidance without conjugation gating and applies the common endpoint analyzer post hoc; *gate-only* enforces conjugation validity without affinity guidance; and *guided* uses both. All four conditions share the same SMILES flow checkpoint and are evaluated on their full valid sets.

Table A1: PDC validity-gate and affinity-guidance ablation. Each cell reports conjugation validity (%) / mean predicted pKd over the full valid set. The pKd predictor also supplies the affinity-guidance score and therefore evaluates the optimized objective.

Length	Configuration	ERBB2	GNRHR	NRP1
80	Unconditional	22.8 / 6.12	22.8 / 4.48	6.1 / 5.20
	Staged	22.7 / 7.70	18.3 / 8.09	23.7 / 8.68
	Gate-only	100.0 / 7.33	100.0 / 6.59	100.0 / 7.88
	Guided	**100.0 / 7.69**	**100.0 / 7.55**	**100.0 / 8.58**
100	Unconditional	24.5 / 6.22	24.5 / 4.93	5.0 / 5.47
	Staged	36.7 / 7.92	34.7 / 8.15	28.1 / 8.65
	Gate-only	100.0 / 7.52	100.0 / 7.01	100.0 / 8.01
	Guided	**100.0 / 7.86**	**100.0 / 8.07**	**100.0 / 8.67**

Across both token budgets and all targets, gating increased conjugation validity from 5.0-36.7% to 100%. Comparing gate-only with guided generation showed that affinity guidance contributed approximately 0.3-1.1 predicted pKd after restricting sampling to the feasible region. The staged condition attained similar endpoint pKd among the small fraction of samples that assembled. Guided generation increased conjugation validity by 2.7-5.5-fold relative to staged generation, with mean predicted binding scores within 0.54 pKd units. These results separate the contributions: the hard gate removed assembly failures, and affinity guidance recovered high predicted binding within the valid manifold.

A.2 PDC Prior Domain Gap and Runtime-Normalized Yield

Choice of peptide prior. We retained the broad pretrained PepReDi prior for PDCs. PDCdb contains only 393 unique canonical peptides within the 6-49 residue pretraining range and 195 unique peptide SMILES, making PDC-specific fine-tuning susceptible to memorization. Only 5 of the 393 canonical PDCdb peptides occur in the pretraining corpus. We therefore measured zero-shot negative log likelihood (NLL) under the base PepReDi model and used 5,000 sequence-level bootstrap samples for confidence intervals (Table A2).

Table A2: Zero-shot NLL under base PepReDi. SEM is computed across sequences; confidence intervals for the NLL difference use 5,000 sequence-level bootstrap samples.

Group	n	Mean NLL (SEM)	ΔNLL	95% CI
PDCdb, all	393	1.5052 (0.0090)	-0.0172	$[-0.0360, 0.0017]$
PDCdb, CPP-like	149	1.4630 (0.0138)	-0.0594	$[-0.0874, -0.0313]$
PDCdb, non-CPP	244	1.5349 (0.0118)	0.0124	$[-0.0115, 0.0364]$
Pretraining validation	1,414	1.5224 (0.0039)		reference

Neither the complete PDCdb set nor its non-CPP subset differed significantly from held-out pretraining data. The lower NLL of the CPP-like subset may reflect its lower sequence complexity and did not indicate a PDC-wide domain shift. These measurements supported retaining the broader peptide prior without fine-tuning on the small, heterogeneous PDCdb set.

Matched sampling comparison. For PDCs, AReUReDi and CupOFMoCA use identical base-flow checkpoints, so their comparison isolates sampling. AReUReDi uses annealed guidance, Tchebycheff scalarization, locally balanced proposals, and Metropolis-Hastings acceptance; CupOFMoCA directly reweights each local proposal with the functional score and validity mask. Across the 12 target-representation blocks formed by SMILES-80, SMILES-100, AA, and UAA generation on ERBB2, GNRHR, and NRP1, CupOFMoCA-Guided achieved the best runtime-normalized valid yield in 8 blocks; AReUReDi-Guided achieved the best yield in none (Table A3). CupOFMoCA reduced runtime by 30-60% in multiple settings and preserved endpoint affinity.

For the runtime-normalized comparison, each representation was evaluated using the same implementation employed to generate its reported samples. In particular, AA- and UAA-tokenized CupOFMoCA-Guided runs used the $\kappa = 8$ cascaded validity evaluation described in Appendix G.2.2, whereas SMILES-tokenized runs evaluated the full Top-K candidate set.

Table A3: Number of wins across 12 target-representation blocks. Valid yield/GPU-hour counts chemically valid conjugates produced per unit compute.

Method	Best mean affinity	Best valid yield/GPU-hour
AReUReDi-Staged	3	1
CupOFMoCA-Staged	**8**	3
AReUReDi-Guided	1	0
CupOFMoCA-Guided	0	**8**

A.3 Conjugation Validity and Binding Affinity

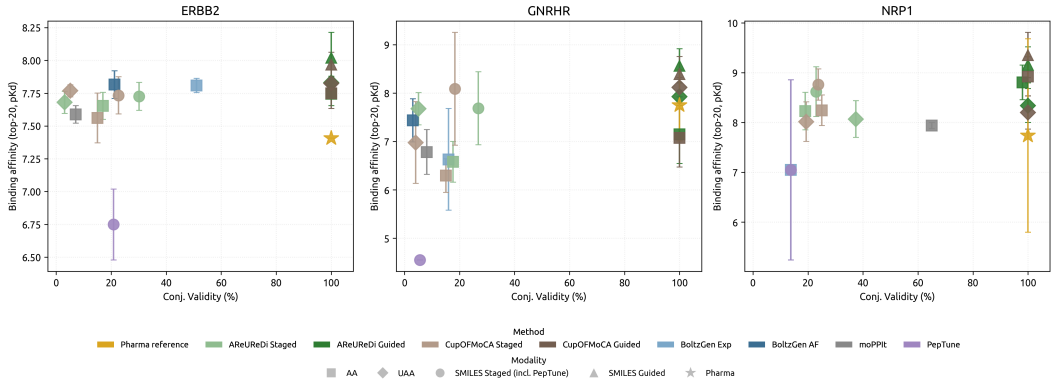

Figure A1: **Joint validity-binding tradeoff for de novo PDC generation across three receptor targets.** Each point shows one method's conjugate validity (x-axis, fraction of valid peptides that yield a valid conjugate) against the mean predicted binding affinity of its top-20 PDCs (y-axis, pKd, error bars: ± 1 std). Method family is encoded by colour (CupOFMoCA Staged: tan; CupOFMoCA Guided: dark taupe; AReUReDi Staged/Guided: light/dark green; baselines: blue/grey/lavender; pharmaceutical reference: gold). Marker shape encodes the input/output representation (square: amino-acid sequence model; diamond: unnatural-AA model; circle: SMILES staged generation, including PepTune; triangle: SMILES guided/coupled generation; star: pharma reference). Guided methods achieved high conjugation validity across all three targets, with representation-dependent predicted binding scores. SMILES-guided samples are triangles; guided AA and UAA samples are squares and diamonds, respectively. Pharma validity is set to 100% by definition (well-formed reference compounds).

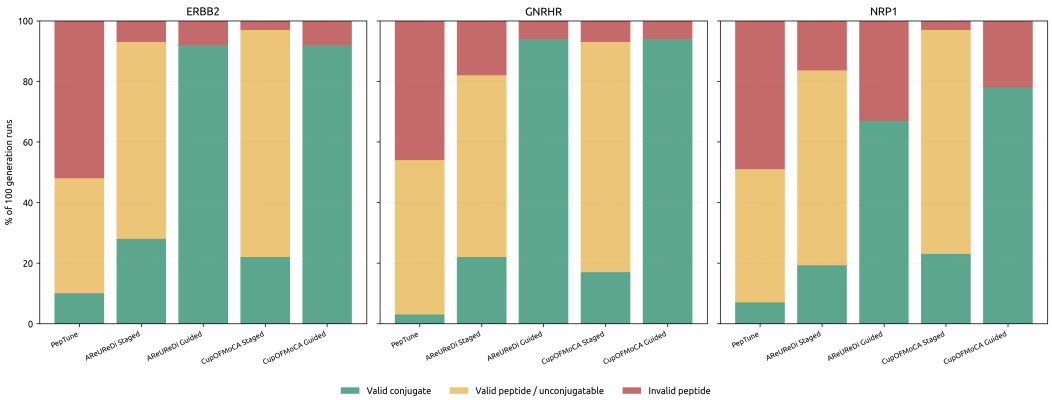

Figure A2: **Decomposition of 100 generation runs by outcome class for SMILES-based methods at peptide length 80.** For each method × target, the bar splits the 100 runs into three categories: invalid peptide (red, generation failed at the SMILES level), valid peptide that fails post-hoc conjugation (orange, lacks the chemical features required for linker attachment), and valid PDC (green, full success). Coupled methods (AReUReDi Guided, CupOFMoCA Guided) eliminate the orange band almost entirely on ERBB2 and GNRHR by generating conjugation-compatible peptides directly, recovering 92-94% valid PDCs vs. 17-28% for staged baselines and 3-10% for PepTune. On NRP1, coupled methods yielded 67-78% valid PDCs across all generation attempts, compared with 7-23% for staged methods.

Table A4: Top-20 *de novo* PDCs across three receptor targets generated by SMILES-based methods (peptide length: 100 SMILES tokens). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). Values are mean (std). **Bold**: best per column; underline: second-best.

	ERBB2		**GNRHR**		**NRP1**	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
PepTune	17.9%	6.82 (0.50)	17.2%	5.85 (1.25)	11.3%	7.25(1.11)
AReUReDi						
Staged	40.9%	7.94 (0.16)	39.1%	**9.15 (0.35)**	18.8%	8.75 (0.64)
Guided	**100.0%**	7.97 (0.20)	**100.0%**	8.86 (0.32)	**100.0%**	**9.23 (0.24)**
CupOFMoCA						
Staged	36.7%	8.09 (0.17)	34.7%	8.66 (0.32)	28.1%	8.85 (0.38)
Guided	**100.0%**	**8.24 (0.14)**	**100.0%**	9.01 (0.36)	**100.0%**	9.11 (0.19)

Table A5: Full-valid-set *de novo* PDCs across three receptor targets generated by AA-based methods (peptide length: 10 AAs). Metrics are computed over all valid conjugates from each method (not filtered to top-20). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). Values are mean (std). **Bold**: best per column; underline: second-best.

	ERBB2		GNRHR		NRP1	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
moPPIt	7.1%	7.59 (0.07)	8.1%	6.78 (0.46)	65.0%	7.73 (0.34)
BoltzGen						
Exp	51.0%	7.64 (0.22)	16.0%	6.63 (1.05)	-	-
AF	21.2%	**7.80 (0.13)**	3.0%	7.44 (0.45)	13.7%	7.05 (1.81)
AReUReDi						
Staged (AA)	17.0%	7.65 (0.10)	17.6%	6.58 (0.42)	19.0%	8.23 (0.38)
Guided (AA)	**100.0%**	7.57 (0.16)	**100.0%**	6.48 (0.49)	98.0%	8.24 (0.37)
Staged (UAA)	3.1%	7.68 (0.08)	5.1%	**7.68 (0.33)**	37.4%	7.92 (0.32)
Guided (UAA)	**100.0%**	7.72 (0.08)	**100.0%**	7.39 (0.47)	**100.0%**	7.94 (0.27)
CupOFMoCA						
Staged (AA)	15.0%	7.56 (0.19)	15.0%	6.30 (0.35)	25.0%	8.16 (0.32)
Guided (AA)	**100.0%**	7.56 (0.14)	**100.0%**	6.40 (0.50)	**100.0%**	**8.28 (0.42)**
Staged (UAA)	5.1%	7.77 (0.05)	4.0%	6.98 (0.84)	19.4%	8.02 (0.40)
Guided (UAA)	**100.0%**	7.71 (0.09)	**100.0%**	7.57 (0.45)	**100.0%**	7.95 (0.22)

Table A6: Full-valid-set *de novo* PDCs across three receptor targets generated by SMILES-based methods (peptide length: 80 SMILES tokens). Metrics are computed over all valid conjugates from each method (not filtered to top-20). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). Values are mean (std). **Bold**: best per column; underline: second-best.

	ERBB2		GNRHR		NRP1	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
PepTune	20.8%	6.75 (0.27)	5.6%	4.55	13.7%	7.05 (1.81)
AReUReDi						
Staged	30.1%	7.65 (0.16)	26.8%	7.54 (0.86)	23.0%	8.62 (0.50)
Guided	**100.0%**	7.68 (0.25)	**100.0%**	7.64 (0.68)	**100.0%**	8.59 (0.50)
CupOFMoCA						
Staged	22.7%	**7.70 (0.17)**	18.3%	**8.09 (1.17)**	23.7%	**8.68 (0.37)**
Guided	**100.0%**	7.69 (0.22)	**100.0%**	7.55 (0.68)	**100.0%**	8.58 (0.64)

Table A7: Full-valid-set *de novo* PDCs across three receptor targets generated by SMILES-based methods (peptide length: 100 SMILES tokens). Metrics are computed over all valid conjugates from each method (not filtered to top-20). Each conjugate comprises a peptide linked to a target-specific cytotoxic payload via the corresponding linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). Values are mean (std). **Bold**: best per column; underline: second-best.

	ERBB2		GNRHR		NRP1	
Method	Conj. Val.	Binding	Conj. Val.	Binding	Conj. Val.	Binding
PepTune	17.9%	6.82 (0.50)	17.2%	5.85 (1.25)	11.3%	7.25 (1.11)
AReUReDi						
Staged	40.9%	7.79 (0.21)	39.1%	**8.57 (0.93)**	18.8%	**8.75 (0.64)**
Guided	**100.0%**	7.69 (0.21)	**100.0%**	7.95 (0.63)	**100.0%**	8.66 (0.66)
CupOFMoCA						
Staged	36.7%	**7.92 (0.25)**	34.7%	8.15 (0.74)	28.1%	8.65 (0.53)
Guided	**100.0%**	7.86 (0.27)	**100.0%**	8.07 (0.74)	**100.0%**	8.67 (0.47)

A.4 Other molecular properties

We compared linker-payload mass increments and peptide masses to interpret the target-specific property distributions (Tables A8 and A9; Figure A3).

Table A8: Per-attachment mass added by each receptor-specific linker-payload is conserved between PDCdb references and generated PDCs, confirming identical conjugation chemistry. $\Delta\mathrm{MW}$ is computed as conjugate MW $-$ peptide MW. Theoretical values: SSCpt (amide, $-\mathrm{H_2O}$) ester/carbonate, AoaDau (amide, $-\mathrm{H_2O}$), DSPASN38 (disulfide, $-2\mathrm{H}$).

Target	Linker-payload (reaction)	Theoretical $\Delta\mathrm{MW}$	Pharma $\Delta\mathrm{MW}$ (n)	Generated $\Delta\mathrm{MW}$ (n)
ERBB2	SSCpt, ester/carbonate (amide)	537.10 / 553.10	538.8 (1)	538.6 ± 0.64 (92)
GNRHR	AoaDau (amide)	582.18	583.5 ± 0.32 (4)	583.7 ± 0.48 (94)
NRP1	DSPASN38, pos10/pos20 (disulfide)	478.12	479.0 ± 0.10 (4)	479.6 ± 0.41 (78)

Table A9: Peptide-length comparison between PDCdb references and generated PDCs. With the linker-payload mass held constant (Table A8), the apparent offsets in Figure A3 reduce to peptide-mass differences. Generated peptides are sampled at a fixed budget of 80 SMILES tokens and frequently use non-canonical residues, so a residue count is not directly comparable to the canonical pharma peptides; we report peptide MW (Da) for both sides and `seq. length` only for the pharma references.

Target	Pharma reference		Generated (80 SMILES tokens)	Direction of offset
	peptide MW (Da)	seq. length	peptide MW (Da)	
ERBB2	1852 (n=1)	18	1531 (1205-1986)	generated *lighter* than pharma
GNRHR	1559 (1072-2300)	13.2 (8-21)	1544 (1218-1907)	matched (within $\sim$15 Da)
NRP1	744 (577-974)	6.8 (5-10)	1396 (1123-1671)	generated *heavier* than pharma

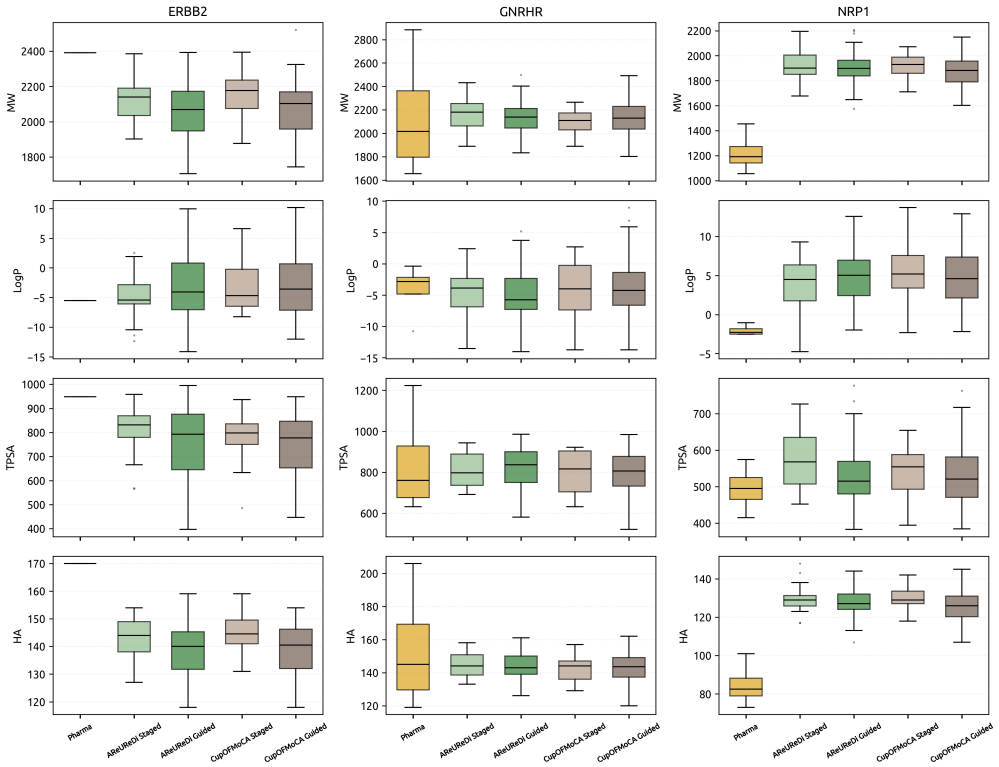

Figure A3: **Molecular property distributions of generated PDCs vs. PDCdb pharmaceutical references at peptide length 80 SMILES tokens.** Box plots show the distributions of molecular weight (MW), lipophilicity (LogP), topological polar surface area (TPSA, Å^2), and heavy-atom count (HA) over all valid conjugates from each SMILES-based method (n=67−94 per cell for AReUReDi/CupOFMoCA Guided; n=17−28 for staged variants), compared against PDCdb reference compounds (gold, n=1 for ERBB2; n=4 for GNRHR; n=4 for NRP1). For ERBB2, the single available reference appears as a degenerate line. All methods use the same receptor-specific linker-payload as the references (ERBB2: SSCpt; GNRHR: AoaDau; NRP1: DSPASN38), and we verified that the per-attachment ΔMW (conjugate − peptide) is conserved between pharma and generated sets within $\sim$1 Da, confirming identical conjugation chemistry on both sides. Differences between gold and generated boxes therefore reflect peptide-length differences rather than chemistry: the GNRHR pharma references are 8-21-residue peptides (mean peptide MW $\approx$1.56 kDa) and overlap the generated distribution; the single ERBB2 reference is an 18-residue peptide ($\approx$1.85 kDa) and sits above the generations; the NRP1 references are short 5-9-residue C-end-rule binders (e.g. CRGDK; mean peptide MW $\approx$0.74 kDa) and sit below. Generated distributions across methods are highly overlapping, indicating that coupled / guided generation does not collapse the chemistry to a narrow region of property space.

Table A10: Full-valid-set molecular properties for AA-based methods on target ERBB2 (peptide length: 10 AAs). All conjugates comprise a peptide linked to a target-specific cytotoxic payload via the corresponding linker). Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). MW: molecular weight; HBD/HBA: hydrogen-bond donors/acceptors. Values are mean (std).

		BoltzGen		*AReUReDi*				*CupOFMoCA*			
Property	**moPPIt**	Exp	AF	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)
Conj. Validity	7.1%	51.0%	21.2%	17.0%	100.0%	3.1%	100.0%	15.0%	100.0%	5.1%	100.0%
Binding affinity	7.59 (0.07)	7.64 (0.22)	7.80 (0.13)	7.65 (0.10)	7.57 (0.16)	7.68 (0.08)	7.72 (0.08)	7.56 (0.19)	7.56 (0.14)	7.77 (0.05)	7.71 (0.09)
MW (Da)	1797.19 (94.33)	1862.70 (262.50)	1652.95 (339.07)	1958.14 (104.89)	1987.75 (110.63)	2460.77 (288.16)	2558.94 (454.66)	2062.10 (90.96)	2037.29 (98.27)	2238.06 (380.48)	2465.91 (404.95)
LogP	2.23 (1.93)	−0.84 (2.37)	−1.03 (2.89)	0.93 (1.94)	0.53 (2.54)	8.93 (2.53)	8.96 (5.73)	0.89 (2.23)	0.52 (1.99)	4.59 (3.27)	6.65 (5.25)
TPSA (Å^2)	556.56 (30.58)	663.67 (120.94)	571.81 (171.79)	648.83 (78.14)	671.25 (83.64)	668.58 (64.45)	723.24 (97.92)	708.49 (84.67)	698.27 (78.00)	670.06 (82.08)	724.58 (111.19)
HBD	17.14 (1.35)	19.27 (4.27)	16.33 (6.55)	19.12 (3.97)	20.54 (3.97)	21.67 (2.08)	21.07 (3.56)	23.07 (4.04)	22.05 (4.00)	20.20 (2.17)	22.23 (3.83)
HBA	26.29 (1.38)	28.33 (3.83)	25.43 (4.60)	27.59 (1.42)	27.86 (1.41)	32.67 (1.15)	34.49 (4.98)	28.27 (1.39)	27.90 (1.38)	32.20 (4.32)	33.87 (4.43)
Rotatable bonds	49.86 (3.34)	54.39 (9.50)	45.14 (12.97)	50.65 (5.20)	51.86 (5.25)	70.67 (5.03)	73.06 (17.43)	55.67 (4.70)	53.52 (4.94)	60.40 (13.16)	69.56 (15.28)
Heavy atoms	124.57 (8.04)	130.10 (18.46)	115.62 (23.77)	138.24 (8.00)	140.21 (8.35)	170.33 (22.50)	178.37 (32.32)	145.33 (7.08)	144.01 (7.26)	155.80 (25.87)	172.12 (28.52)

Table A11: Full-valid-set molecular properties for AA-based methods on target GNRHR (peptide length: 10 AAs). All conjugates comprise a peptide linked to a target-specific cytotoxic payload via the corresponding linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). MW: molecular weight; HBD/HBA: hydrogen-bond donors/acceptors. Values are mean (std).

		BoltzGen		*AReUReDi*				*CupOFMoCA*			
Property	**moPPIt**	Exp	AF	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)
Conj. Validity	8.1%	16.0%	3.0%	17.6%	100.0%	5.1%	100.0%	15.0%	100.0%	4.0%	100.0%
Binding affinity	6.78 (0.46)	6.63 (1.05)	7.44 (0.45)	6.58 (0.42)	6.48 (0.49)	7.68 (0.33)	7.39 (0.47)	6.30 (0.35)	6.40 (0.50)	6.98 (0.84)	7.57 (0.45)
MW (Da)	1807.91 (122.97)	1914.95 (267.16)	1629.48 (178.58)	2057.01 (112.57)	2044.31 (106.20)	2431.32 (446.81)	2506.25 (395.02)	2081.44 (99.03)	2074.30 (110.34)	2852.63 (370.48)	2650.75 (445.82)
LogP	−0.08 (3.66)	−3.07 (3.00)	0.36 (1.88)	−2.58 (1.74)	−2.51 (1.40)	2.62 (3.27)	4.77 (6.72)	−1.87 (1.42)	−2.39 (1.79)	8.73 (11.32)	7.00 (8.38)
TPSA (Å^2)	630.78 (69.49)	712.12 (138.16)	582.29 (78.16)	794.03 (91.01)	782.39 (73.53)	835.99 (164.41)	806.31 (98.75)	787.20 (74.23)	795.54 (86.98)	836.49 (137.88)	802.97 (104.47)
HBD	20.62 (2.26)	21.62 (4.79)	17.00 (3.00)	26.56 (4.57)	25.80 (3.76)	25.00 (4.30)	25.44 (3.73)	26.07 (4.04)	26.75 (4.62)	26.25 (4.65)	26.01 (4.21)
HBA	28.25 (1.67)	29.62 (4.08)	25.00 (2.00)	30.61 (1.14)	30.29 (1.15)	35.80 (6.69)	35.72 (4.05)	30.27 (0.88)	30.24 (1.35)	39.25 (1.71)	36.58 (3.84)
Rotatable bonds	46.25 (4.40)	49.19 (9.16)	41.00 (7.81)	52.22 (5.19)	51.00 (5.36)	64.80 (16.45)	67.25 (13.60)	52.13 (6.00)	51.94 (5.76)	77.75 (12.61)	72.08 (15.56)
Heavy atoms	128.00 (9.64)	136.12 (19.05)	116.33 (12.86)	146.56 (8.40)	145.83 (7.89)	172.20 (31.70)	177.40 (27.69)	148.73 (7.15)	148.08 (8.15)	201.00 (25.44)	187.29 (31.77)

Table A12: Full-valid-set molecular properties for AA-based methods on target NRP1 (peptide length: 10 AAs). All conjugates comprise a peptide linked to SN-38 payload via Disulfanylpropanoic Acid linker. Conjugation Validity: fraction of valid peptides yielding chemically valid conjugates. Binding affinity: predicted peptide-receptor binding score (pKd). MW: molecular weight; HBD/HBA: hydrogen-bond donors/acceptors. Values are mean (std).

			AReUReDi				*CupOFMoCA*			
Property	**moPPIt**	**BoltzGen AF**	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)	Staged (AA)	Guided (AA)	Staged (UAA)	Guided (UAA)
Conj. Validity	65.0%	13.7%	19.0%	98.0%	37.4%	100.0%	25.0%	100.0%	19.4%	100.0%
Binding affinity	7.73 (0.34)	7.05 (1.81)	8.23 (0.38)	8.24 (0.37)	7.92 (0.32)	7.94 (0.27)	8.16 (0.32)	8.28 (0.42)	8.02 (0.40)	7.95 (0.22)
MW (Da)	1754.71 (96.21)	1680.54 (141.90)	1838.20 (115.56)	1842.31 (82.55)	2431.86 (271.82)	2325.85 (252.47)	1854.46 (94.83)	1859.13 (87.62)	2269.39 (317.96)	2244.83 (296.12)
LogP	4.11 (2.09)	2.22 (5.38)	2.87 (2.07)	3.00 (2.07)	10.74 (4.53)	8.75 (3.94)	3.73 (2.07)	3.40 (2.16)	10.37 (4.20)	8.23 (4.44)
TPSA (Å^2)	525.45 (55.18)	518.17 (130.08)	566.37 (70.85)	566.61 (55.68)	626.38 (64.34)	625.54 (78.18)	556.44 (65.69)	566.99 (72.48)	569.27 (61.28)	596.79 (61.02)
HBD	16.12 (2.68)	14.71 (5.09)	17.05 (3.73)	16.53 (2.68)	17.35 (1.77)	17.40 (2.50)	16.40 (3.01)	16.58 (3.41)	16.37 (1.86)	17.00 (2.04)
HBA	24.94 (1.92)	24.86 (4.30)	25.42 (1.61)	25.50 (1.29)	32.62 (3.34)	31.85 (3.66)	25.04 (1.02)	25.24 (1.31)	29.74 (3.53)	30.82 (3.65)
Rotatable bonds	44.29 (2.87)	41.14 (8.32)	46.05 (6.17)	46.05 (4.52)	68.92 (12.30)	63.57 (10.73)	46.64 (4.56)	46.44 (4.49)	63.37 (12.48)	61.18 (11.97)
Heavy atoms	123.40 (7.59)	117.43 (9.93)	129.37 (8.23)	129.64 (6.14)	169.35 (19.60)	162.31 (17.60)	130.64 (7.04)	131.02 (6.54)	157.95 (22.13)	156.61 (20.56)

Table A13: Full-valid-set SMILES-based methods, target ERBB2 (peptide length 80 SMILES tokens).

Property	**PepTune**	**AReUReDi Staged**	**AReUReDi Guided**	**CupOFMoCA Staged**	**CupOFMoCA Guided**
Conj. Validity	20.8%	30.1%	100.0%	22.7%	100.0%
Binding affinity	6.75 (0.27)	7.65 (0.16)	7.68 (0.25)	7.70 (0.17)	7.69 (0.22)
MW (Da)	1875.21 (110.72)	2117.46 (121.51)	2062.85 (150.76)	2159.62 (125.87)	2069.22 (156.24)
LogP	−1.32 (5.03)	−4.78 (3.55)	−3.25 (5.32)	−3.13 (4.08)	−3.07 (5.04)
TPSA (Å^2)	665.82 (97.57)	807.02 (101.27)	750.90 (142.53)	779.71 (92.59)	750.40 (131.44)
HBD	18.60 (4.03)	25.64 (3.64)	25.01 (5.26)	25.05 (3.36)	24.35 (4.74)
HBA	28.80 (3.55)	43.39 (3.85)	42.76 (5.77)	44.32 (4.68)	42.59 (5.70)
Rotatable bonds	53.80 (6.51)	72.54 (7.94)	71.77 (8.38)	75.73 (8.02)	71.07 (7.44)
Heavy atoms	130.40 (6.82)	142.96 (7.87)	138.96 (9.90)	145.00 (7.66)	138.72 (9.69)

Table A14: Full-valid-set SMILES-based methods, target GNRHR (peptide length 80 SMILES tokens).

Property	**PepTune**	**AReUReDi Staged**	**AReUReDi Guided**	**CupOFMoCA Staged**	**CupOFMoCA Guided**
Conj. Validity	5.6%	26.8%	100.0%	18.3%	100.0%
Binding affinity	4.55	7.54 (0.86)	7.64 (0.68)	8.09 (1.17)	7.55 (0.68)
MW (Da)	1857.06 (0.00)	2161.09 (140.02)	2136.08 (132.55)	2100.42 (90.95)	2128.04 (142.41)
LogP	−4.53	−4.38 (3.92)	−5.03 (4.20)	−3.92 (4.57)	−3.84 (4.02)
TPSA (Å^2)	777.09	814.26 (87.94)	821.02 (98.05)	791.03 (105.28)	795.98 (105.34)
HBD	27.00	30.45 (3.73)	29.63 (3.96)	27.76 (3.70)	28.76 (4.21)
HBA	32.00	46.41 (4.97)	45.27 (4.49)	44.76 (4.32)	45.69 (4.79)
Rotatable bonds	51.00	73.27 (6.51)	69.04 (7.75)	70.00 (8.06)	70.62 (7.48)
Heavy atoms	132.00	144.91 (7.62)	143.71 (8.37)	142.53 (6.98)	142.67 (8.80)

Table A15: Full-valid-set SMILES-based methods, target NRP1 (peptide length 80 SMILES tokens).

Property	**PepTune**	**AReUReDi Staged**	**AReUReDi Guided**	**CupOFMoCA Staged**	**CupOFMoCA Guided**
Conj. Validity	13.7%	23.0%	100.0%	23.7%	100.0%
Binding affinity	7.05 (1.81)	8.62 (0.50)	8.59 (0.50)	8.68 (0.37)	8.58 (0.64)
MW (Da)	1680.54 (141.90)	1928.14 (132.97)	1907.99 (115.10)	1925.42 (95.30)	1875.63 (124.12)
LogP	2.22 (5.38)	3.93 (3.67)	4.98 (3.32)	5.09 (3.89)	4.88 (3.55)
TPSA (Å^2)	518.17 (130.08)	576.12 (85.53)	528.84 (72.29)	543.50 (68.90)	528.09 (81.09)
HBD	14.71 (5.09)	19.45 (2.86)	17.40 (2.96)	17.61 (4.05)	17.82 (3.58)
HBA	24.86 (4.30)	37.40 (5.09)	36.01 (3.26)	36.22 (4.32)	36.21 (3.81)
Rotatable bonds	41.14 (8.32)	69.85 (8.36)	67.21 (8.28)	68.09 (9.02)	67.00 (8.32)
Heavy atoms	117.43 (9.93)	129.35 (7.59)	128.22 (6.81)	129.78 (6.13)	125.91 (7.59)

Table A16: Full-valid-set SMILES-based methods, target ERBB2 (peptide length 100 SMILES tokens).

Property	PepTune	AReUReDi Staged	AReUReDi Guided	CupOFMoCA Staged	CupOFMoCA Guided
Conj. Validity	17.9%	40.9%	100.0%	36.7%	100.0%
Binding affinity	6.82 (0.50)	7.79 (0.21)	7.69 (0.21)	7.92 (0.25)	7.86 (0.27)
MW (Da)	2135.35 (135.67)	2458.65 (158.09)	2422.54 (176.41)	2513.95 (147.00)	2436.57 (171.09)
LogP	0.92 (3.59)	−3.76 (4.20)	−3.98 (4.88)	−3.61 (5.23)	−3.10 (5.44)
TPSA (Å^2)	674.01 (54.77)	905.03 (103.73)	887.69 (136.12)	913.99 (110.20)	875.13 (131.86)
HBD	20.80 (3.56)	28.56 (4.04)	29.03 (4.76)	28.17 (5.85)	28.34 (5.09)
HBA	32.80 (4.55)	49.28 (4.67)	48.36 (5.17)	50.69 (4.57)	49.11 (5.44)
Rotatable bonds	59.20 (7.53)	86.14 (8.64)	85.29 (8.58)	88.86 (8.25)	86.35 (8.95)
Heavy atoms	147.60 (8.14)	164.97 (8.48)	163.27 (11.44)	167.78 (7.57)	163.54 (10.59)

Table A17: Full-valid-set SMILES-based methods, target GNRHR (peptide length 100 SMILES tokens).

Property	PepTune	AReUReDi Staged	AReUReDi Guided	CupOFMoCA Staged	CupOFMoCA Guided
Conj. Validity	17.2%	39.1%	100.0%	34.7%	100.0%
Binding affinity	5.85 (1.25)	8.57 (0.93)	7.95 (0.63)	8.15 (0.74)	8.07 (0.74)
MW (Da)	2180.29 (135.67)	2534.01 (131.60)	2420.67 (220.22)	2544.61 (111.73)	2536.91 (143.44)
LogP	−0.98 (3.59)	−5.39 (3.84)	−6.13 (4.31)	−3.96 (4.25)	−4.81 (4.13)
TPSA (Å^2)	750.86 (54.77)	981.23 (93.21)	953.53 (119.62)	952.33 (119.41)	962.22 (105.10)
HBD	24.80 (3.56)	33.76 (4.20)	32.85 (4.86)	32.76 (4.34)	32.96 (4.70)
HBA	34.80 (4.55)	50.47 (4.07)	50.31 (5.42)	50.79 (4.66)	52.15 (4.75)
Rotatable bonds	57.20 (7.53)	84.97 (8.64)	79.82 (10.79)	87.47 (7.40)	85.90 (8.97)
Heavy atoms	152.60 (8.14)	171.24 (8.41)	163.59 (15.01)	172.12 (8.03)	171.00 (9.69)

Table A18: Full-valid-set SMILES-based methods, target NRP1 (peptide length 100 SMILES tokens).

Property	PepTune	AReUReDi Staged	AReUReDi Guided	CupOFMoCA Staged	CupOFMoCA Guided
Conj. Validity	11.3%	18.8%	100.0%	28.1%	100.0%
Binding affinity	7.25(1.11)	8.75 (0.64)	8.66 (0.66)	8.65 (0.53)	8.67 (0.47)
MW (Da)	2001.96(143.32)	2270.30 (179.81)	2327.55 (147.16)	2243.58 (171.81)	2285.06 (149.25)
LogP	0.04 (2.32)	4.29 (4.16)	3.20 (5.86)	4.36 (4.79)	2.85 (5.81)
683.09 (80.04)	698.83 (110.29)	731.19 (147.33)	684.12 (140.35)	711.35 (153.47)	
HBD	19.83 (1.72)	22.50 (4.10)	24.25 (5.10)	22.26 (4.74)	23.29 (5.64)
HBA	29.83 (3.60)	41.69 (4.44)	44.43 (5.40)	41.30 (5.43)	43.68 (5.49)
Rotatable bonds	54.00 (10.30)	81.69 (7.60)	85.64 (8.87)	78.70 (9.24)	82.60 (9.74)
Heavy atoms	140.50 (9.89)	154.00 (12.28)	156.71 (10.30)	151.30 (10.42)	153.02 (10.34)

A.5 Variable Linker Generation

Table A19: Molecular properties of PDCs generated with variable vs. fixed linkers (GNRHR-Dau: fixed Aoa; ERBB2-Cpt: fixed SS) for the same receptor-payload pairs. All results are from 100 generation runs at peptide length 80 SMILES tokens. Values shown as mean (std).

Property	**GNRHR-Dau Coupled**	**GNRHR-Dau Var. Linker**	**ERBB2-Cpt Coupled**	**ERBB2-Cpt Var. Linker**
Conj. validity	100%	100%	100%	100%
Binding affinity	7.55 (0.68)	7.67 (0.96)	7.69 (0.22)	7.73 (0.23)
MW (Da)	2128 (142)	2136 (178)	2069 (156)	2090 (149)
LogP	−3.84 (4.02)	−3.58 (3.92)	−3.07 (5.04)	−2.82 (4.98)
TPSA (Å^2)	796.0 (105.3)	796.6 (102.1)	750.4 (131.4)	760.9 (137.8)
HBD	28.8 (4.2)	28.9 (4.2)	24.4 (4.7)	24.2 (4.6)
HBA	45.7 (4.8)	45.1 (3.8)	42.6 (5.7)	43.2 (5.6)
Rotatable bonds	70.6 (7.5)	70.5 (9.3)	71.1 (7.4)	71.7 (7.1)
Heavy atoms	142.7 (8.8)	144.2 (12.0)	138.7 (9.7)	141.0 (10.0)

Table A20: Linker type diversity for the two receptor-payload pairs. **DB** refers to the PDCdb database distribution of linker types. **Joint** stands for generation with randomly sampled linker type drawn from the implemented linker types' frequencies in PDCdb.

(a) GNRHR receptor, Daunorubicin payload.

Linker	**DB Count**	**DB %**	**Joint Count**	**Joint %**
Aoa	21	77.8%	78	84.8%
Glutaryl	2	7.4%	6	6.5%
ValAla	2	7.4%	4	4.3%
ValCit	2	7.4%	4	4.3%

(b) ERBB2 receptor, Camptothecin payload.

Linker	**DB Count**	**DB %**	**Joint Count**	**Joint %**
DSEC	5	18.5%	33	35.9%
SS	7	25.9%	23	25.0%
MalHex	5	18.5%	21	22.8%
DSPA	5	18.5%	15	16.3%

Table A21: Generation diversity metrics across receptors, peptide lengths, and linker configurations. Validity: fraction of generated peptides forming valid conjugates. Uniqueness: fraction of distinct samples among generated. Diversity: average pairwise Tanimoto distance. SNN: similarity to nearest neighbor in training set (lower = more novel).

Receptor	**Variant**	**Length**	**Validity**	**Uniqueness**	**Diversity**	**SNN**
ERBB2	Coupled	80	0.919	1.000	0.844	0.187
		100	0.949	1.000	0.835	0.194
	Var. linker	80	0.919	1.000	0.842	0.188
		100	0.960	1.000	0.835	0.194
GNRHR	Coupled	80	0.939	1.000	0.843	0.180
		100	0.960	1.000	0.832	0.187
	Var. linker	80	0.919	1.000	0.845	0.179
		100	0.949	1.000	0.832	0.189

A.6 Structural Validation

To evaluate the structural quality of designs generated by CupOFMoCA, we performed structure prediction using Boltz-2, the current state-of-the-art molecule-protein complex predictor [Passaro et al., 2025], on receptor-ligand complexes. We compared our generated designs (Coupled) against reference pharmaceutical compounds (Pharma) across four metrics: overall confidence score, interface predicted TM-score (ipTM), complex pLDDT, and interface pLDDT (ipLDDT).

Table A22: Structure prediction quality for top 10 ERBB2-targeting designs (length 80). Comparison of peptides and PDCs binding to ERBB2 receptor, with metrics measured on the receptor-ligand complex using Boltz-2. Pharma represents reference pharmaceutical designs, while Coupled represents our generation method. conf. = confidence score, ipTM = interface predicted TM-score, pLDDT = complex pLDDT, ipLDDT = complex ipLDDT.

Metric	Peptide		**Conjugate**	
	Pharma	**Ours**	Pharma	**Ours**
Conf.	0.6653	0.6878	0.6415	0.6971
ipTM	0.7217	0.7571	0.6281	0.8043
pLDDT	0.6511	0.6705	0.6448	0.6703
ipLDDT	0.4344	0.4976	0.4021	0.5023

Table A23: Structure prediction quality for top 10 GNRHR-targeting designs (length 80). Comparison of peptides and PDCs binding to GNRHR receptor, with metrics measured on the receptor-ligand complex using Boltz-2.

Metric	Peptide		**Conjugate**	
	Pharma	**Ours**	Pharma	**Ours**
Conf.	0.8037	0.7764	0.7111	0.7251
ipTM	0.9217	0.8826	0.8581	0.8695
pLDDT	0.7742	0.7498	0.6744	0.6890
ipLDDT	0.6522	0.5827	0.4639	0.4961

Table A24: Structure prediction quality for top 10 NRP1-targeting designs (length 80). Comparison of peptides and PDCs binding to NRP1 receptor, with metrics measured on the receptor-ligand complex using Boltz-2.

Metric	Peptide		**Conjugate**	
	Pharma	**Ours**	Pharma	**Ours**
Conf.	0.7314	0.6545	0.6560	0.6356
ipTM	0.6673	0.3914	0.4359	0.3486
pLDDT	0.7474	0.7203	0.7111	0.7074
ipLDDT	0.6563	0.4421	0.4547	0.4036

As shown in Tables A22, A23, and A24, CupOFMoCA produces structurally valid receptor-binding designs across three diverse targets, with results that vary by target but reveal a consistent advantage on conjugates relative to peptides. For ERBB2, our method outperforms pharmaceutical references across all four metrics for both peptides and PDCs, with particularly substantial improvements on conjugates: a 28% relative gain in ipTM (0.8043 vs. 0.6281) and 25% in ipLDDT (0.5023 vs. 0.4021). For GNRHR, CupOFMoCA-generated conjugates exceeded the pharmaceutical references on every metric. This conjugate-level gain accompanied lower scores for the corresponding peptide designs, supporting the value of evaluating the assembled molecule. For NRP1, pharmaceutical references retain an advantage on both modalities; however, the gap is markedly narrower for conjugates than for peptides (e.g., a confidence gap of 0.020 for PDCs vs. 0.077 for peptides), echoing the same pattern seen on the other two targets. Taken together, these results suggest that our coupled generation approach is particularly effective for the more challenging task of designing PDCs, where the model must jointly optimize peptide sequence and conjugation chemistry while maintaining favorable binding geometry; closing the remaining gap on harder targets such as NRP1 is a natural direction for future work.

B PROTAC Additional Results

B.1 Generated Samples

We selected three candidates per system by synthetic accessibility from the five highest DegradeMaster-scoring candidates (Figure B1). Their generated linkers are highlighted between the fixed warhead and E3-ligand fragments.

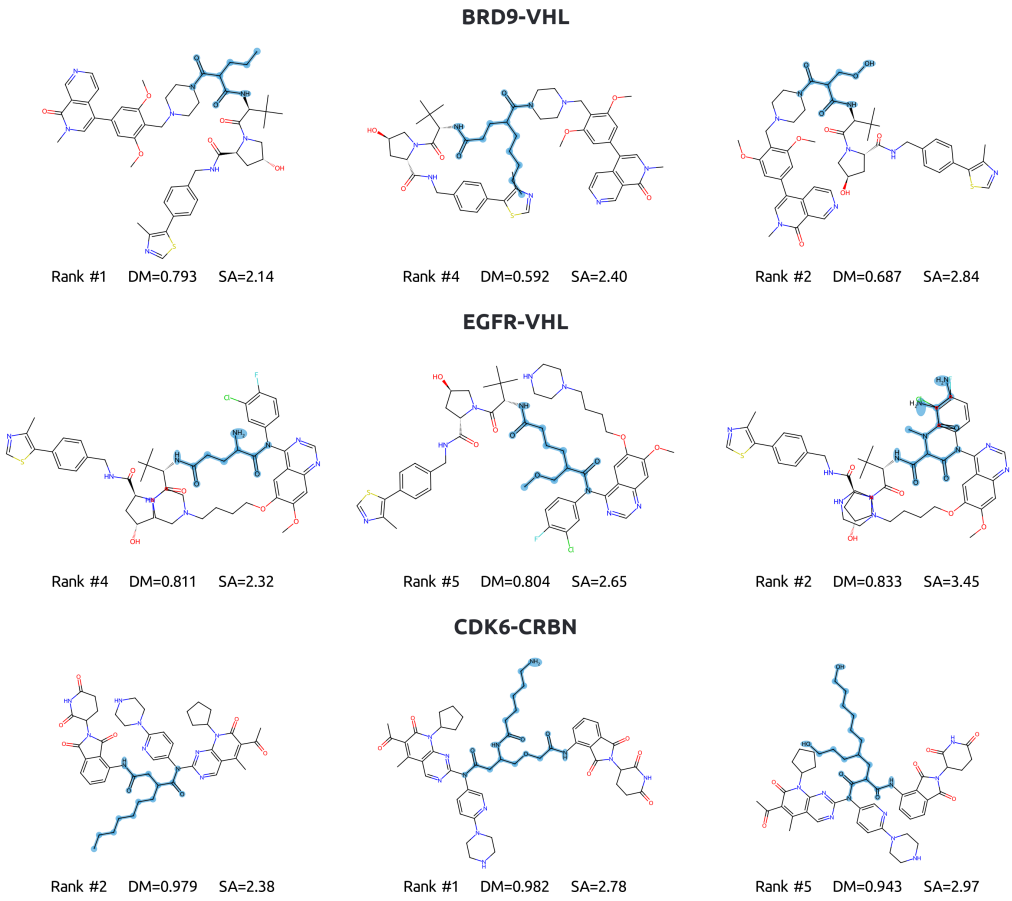

Figure B1: **Top three PROTACs generated for each target.** For each target system (BRD9-VHL, EGFR-VHL, CDK6-CRBN), the three PROTACs with the lowest synthetic-accessibility (SA) score among the top five by DegradeMaster (DM) probability are shown horizontally. Each panel depicts the assembled conjugate with the *de novo*-generated linker highlighted in blue; the warhead and E3-ligand fragments are the fixed inputs. DM denotes the predicted degradation-activity probability (higher is better) and SA the synthetic-accessibility score (lower is better, range 1-10).

B.2 Molecular Properties Baselines Comparison

Table B1: Molecular properties of valid PROTACs across three target systems (EGFR-VHL, BRD9-VHL, CDK6-CRBN). SA ↓, all others ↑. SA and QED reported as mean ± std across molecules. The $\log P$, MW, and RotB rows report the fraction of valid PROTACs falling within PROTAC-DB-derived windows ($\log P \in [2.5, 7.4]$, $\mathrm{MW} \in [771, 1108]$ Da, $\mathrm{RotB} \in [12, 25]$); note that Table H1 uses a wider MW window (700-1200 Da) and is not directly comparable on that row. Our column pools three seeds (286 / 283 / 284 RDKit-valid linkers) under the optimized objective weights; baseline columns are single batches of $N = 100$. Uniqueness is the per-seed mean; pooled over all three seeds it is 92.7 / 92.9 / 91.9%. Dashes (-) indicate 0% assembly validity on that system.

Target	Property	LinkerNet	DiffPROTACs	Ours
EGFR-VHL	SA score	**1.89 ± 0.46**	2.95 ± 0.91	2.76 ± 0.57
	QED	0.487 ± 0.072	0.424 ± 0.133	**0.51 ± 0.13**
	$\log P$ in range	**87.0%**	75.0%	70.3%
	MW in range	**100.0%**	80.0%	71.4%
	RotB in range	**100.0%**	67.5%	84.5%
	Uniqueness	65.2%	**97.5%**	96.2%
BRD9-VHL	SA score	**1.94 ± 0.52**	-	2.76 ± 0.57
	QED	0.457 ± 0.072	-	**0.51 ± 0.12**
	$\log P$ in range	87.5%	-	**99.6%**
	MW in range	**100.0%**	-	90.7%
	RotB in range	**100.0%**	-	95.0%
	Uniqueness	**100.0%**	-	96.8%
CDK6-CRBN	SA score	**1.97 ± 0.44**	2.58 ± 0.79	2.83 ± 0.59
	QED	0.484 ± 0.078	0.513 ± 0.106	**0.51 ± 0.12**
	$\log P$ in range	87.5%	**91.9%**	84.2%
	MW in range	**100.0%**	**100.0%**	**100.0%**
	RotB in range	89.6%	**97.3%**	78.4%
	Uniqueness	77.1%	**100.0%**	96.5%

B.3 Physical Validation by Molecular Dynamics

DegradeMaster is a learned predictor trained on overlapping databases. We therefore evaluated four generated EGFR-VHL candidates, measuring conformational 3D polar surface area in chloroform (a low-dielectric membrane-interior mimic, where lower values indicate stronger permeability-relevant polarity shielding) and the conformational penalty $-kT \log f_{\mathrm{clash-free}}$ in water, where $f_{\mathrm{clash-free}}$ is the fraction of conformers compatible with ternary geometry against EGFR (PDB 4WKQ) and VHL (PDB 6GFY). We included MS39, an experimentally validated cell-active EGFR-VHL degrader sharing our warhead and E3 ligand, as a positive control, and a PEG4-linked congener as a negative control in explicit-solvent molecular dynamics. We ran three independent 50 ns simulations per compound in each solvent following Poongavanam et al. [2022]. Method details are in Appendix B.3.1.

B.3.1 Molecular Dynamics Protocol

We ran all simulations in OpenMM 8.5.2 with OpenFF Sage 2.2.0 and AM1-BCC charges for PROTACs, Amber ff14SB for proteins, and TIP3P water; chloroform boxes were built with Packmol. We used a LangevinMiddle integrator with a 2 fs timestep and 1 ps^{-1} friction, PME electrostatics with a 1.0 nm cutoff, and NPT equilibration. Analysis used RDKit and MDTraj. Solvent-accessible 3D PSA is computed with Bondi radii and a 1.4 Å probe (Shrake-Rupley), matching the PyMOL convention; values are therefore comparable to PyMOL-derived published ranges and read systematically lower than VEGA-derived ones. We validated the 3D PSA implementation on grazoprevir, recovering its published range (median 164 $\mathrm{Å}^2$, with 92% of conformers between 144 and 205 $\mathrm{Å}^2$).

All four generated candidates occupied a conformational regime consistent with the validated degrader. C1 and C2 had lower 3D PSA and lower ternary penalties than MS39. The negative control had the highest values on both metrics, supporting the sensitivity of the analysis to linker-dependent

differences. These simulations measured conformational properties and did not establish cellular degradation. Full target-PROTAC-E3 ternary simulations remain in progress.

B.4 Simulated Compounds

All six compounds share the Gefitinib-derived EGFR warhead and the VH032 VHL binder and differ only in the linker (Table B2). SMILES for all six are in the code repository.

Table B2: Explicit-solvent MD for generated EGFR-VHL PROTACs against experimental controls. MS39 is a validated cell-active EGFR-VHL degrader [Cheng et al., 2020]; PEG4 is a PEG4-linked congener (negative control). Ternary penalty in kcal/mol.

	Linker	**Role**	**3D PSA (Å^2) ↓**	**Penalty ↓**
C1	n-propylsuccinyl	Generated	**140**	**2.2**
C2	allyl succinyl	Generated	**141**	**1.9**
C3	gem-dimethylsuccinate	Generated	178	**2.7**
C4	butylmalonate	Generated	184	**3.3**
MS39	decanedioyl	Positive	178	3.7
PEG4	PEG4	Negative	193	4.1

C Algorithm

C.1 Preliminaries

We introduce the mathematical objects required to define coupled discrete flows on structured molecular string spaces and to formulate objective-guided transport for molecular conjugate assembly. Throughout, $\mathcal{V}$ denotes a finite vocabulary of molecular string tokens (e.g., SMILES tokens), and $\mathcal{V}^L$ denotes the set of molecular strings of length L. In this work, peptides are represented directly as peptide SMILES strings, consistent with prior peptide generative frameworks such as PepTune, TR2-D2, and AReUReDi [Tang et al., 2025a, Chen et al., 2026c, Tang et al., 2025c]. While we present notation for a single molecular string, subsequent sections consider structured states composed of multiple coupled molecular strings corresponding to different components of a conjugate.

C.2 Molecular Strings as States in a Discrete Space

A molecular object is represented as a discrete string state

$$x = (x_1, \dots, x_L) \in \mathcal{V}^L, \tag{7}$$

where each token $x_i \in \mathcal{V}$ corresponds to a valid SMILES symbol. For peptide components, x encodes a peptide SMILES string rather than a residue-level amino-acid sequence.

Transitions between molecular strings occur through local token-level edit operations. We define the edit neighborhood

$$\mathcal{N}(x) := \left\{ x' \in \mathcal{V}^L \,:\, \text{single token edit} \right\}, \tag{8}$$

where a local edit corresponds to a token substitution. Extensions to insertions or deletions are straightforward and compatible with SMILES-based representations, but are omitted here for clarity.

C.3 Reference Dynamics

To define transport in discrete molecular string space, we introduce a simple reference process that captures uninformed local exploration. We consider a continuous-time Markov chain (CTMC) with generator R_0 defined by

$$R_0(x, x') = \begin{cases} \frac{1}{|\mathcal{N}(x)|} & x' \in \mathcal{N}(x), \\ -\sum_{y \neq x} R_0(x, y) & x' = x. \end{cases} \tag{9}$$

This reference dynamics defines a random walk over molecular strings that performs small, unbiased token edits and serves as a neutral baseline relative to which structured transport is learned.

C.4 Discrete Flow

Rather than modeling infinitesimal stochastic dynamics, we focus on learning *discrete flows* that directly transport molecular string states between times. For $0 < t \leq T$, a discrete flow is a function

$$\Phi_\theta^{t \to T} : \mathcal{V}^L \to \mathcal{V}^L, \tag{10}$$

parameterized by θ, which maps a molecular string at time t to a molecular string at time T. These maps are intended to approximate the solution operator of an underlying probability flow, enabling one- or few-step generation without explicit simulation of a continuous-time process.

C.5 Objective Functions

To guide transport toward functionally relevant regions of molecular string space, we assume access to one or more objective functions

$$S : \mathcal{V}^L \to \mathbb{R}, \tag{11}$$

which score molecular strings according to downstream criteria such as peptide-target binding, compatibility with conjugation chemistry, or chemical feasibility. Importantly, these objectives are evaluated on the *current transported molecular state* and do not define a probability density. Instead, they provide directional guidance that biases transport toward desirable configurations.

C.6 Generative Sampling

After training, sampling proceeds by drawing an initial state $X_0 \sim \mu_0$ from a simple molecular-string prior and applying the learned flow:

$$X_t = \Phi_\theta^t(X_0), \qquad 0 < t \leq T. \tag{12}$$

In the multistep regime, we choose a time grid $0 = t_0 < t_1 < \cdots < t_K = T$ and compose maps sequentially. Iterating this procedure from 0 to T yields samples from the learned objective-guided transport, producing valid molecular strings corresponding to peptides or assembled conjugates in a small number of steps.

C.7 Main Algorithm

Algorithm 1 Exponential Tilting with Acceptance Criterion

Input: trained discrete flow Φ_{θ^*}, prior μ_0 on $\mathcal{Z}$
time horizon T, number of steps K
guidance score $S(z)$, strength β
Define time grid $0 = t_0 < t_1 < \cdots < t_K = T$
Sample $Z_0 \sim \mu_0$ s.t. $\mathbf{1}_{\text{valid}}(Z_0) = 1$; $Z^* \leftarrow Z_0$
for $k = 0$ **to** $K - 1$ **do**
  Select random position i
  $p_{k+1} \leftarrow \Phi_{\theta^*}(Z_k)[\cdot, i]$ (predicted distribution at i)
  for each $z' \in \text{TopK}(p_{k+1})$ **do**
    $p'(z') \propto p_{k+1}(z') \exp(\beta S(z'))\, \mathbf{1}_{\text{valid}}(z')$
    (exponential tilting)
  end for
  Sample $z^\dagger \sim p'$
  $Z' \leftarrow Z_k$ with position i set to $z^\dagger$
  if $S(Z') \geq S(Z_k)$ **then**
    (acceptance criterion)
    $Z_{k+1} \leftarrow Z'$
    if $S(Z') > S(Z^*)$ **then**
      $Z^* \leftarrow Z'$
    end if
  else
    $Z_{k+1} \leftarrow Z_k$ (reject, keep current)
  end if
end for
return Z^* (best molecular conjugate found)

D Theoretical Proofs

We provide a minimal theoretical justification for the objective-guided sampling mechanism used in CupOFMoCA. In particular, we show that the exponential tilting rule employed during discrete flow sampling arises as the unique solution to a KL-regularized objective improvement problem under hard validity constraints.

D.1 Notation and Setup

We adopt the notation from the main text. Let $\mathcal{Z}$ denote the finite discrete state space of molecular conjugates. A discrete flow step is represented by a Markov kernel $K(\cdot \mid z)$ on $\mathcal{Z}$, where $K(z' \mid z)$ denotes the probability of proposing state z' given current state z. Deterministic updates are recovered as the special case where $K(\cdot \mid z)$ is a point mass.

Validity is encoded by an indicator function $\mathbf{1}_{\text{valid}} : \mathcal{Z} \to \{0, 1\}$, which specifies whether a candidate state satisfies all chemical and representational constraints. Objective guidance is defined by a score function $S : \mathcal{Z} \to \mathbb{R}$ and a guidance strength $\beta > 0$.

D.2 Objective Guidance as KL-Regularized Exponential Tilting

We show that exponential tilting is the unique optimal way to incorporate objective information into a discrete flow step while remaining close to a base kernel and enforcing hard validity constraints.

Theorem D.1 (Exponential tilting solves KL-regularized objective improvement)**.** *Fix a base kernel $K(\cdot \mid z)$ on $\mathcal{Z}$ and a score function $S : \mathcal{Z} \to \mathbb{R}$. For a fixed input state z, consider the optimization problem*

$$\max_{Q} \left\{ \mathbb{E}_{Z' \sim Q}[S(Z')] - \frac{1}{\beta} \operatorname{KL}(Q \,\|\, K(\cdot \mid z)) \right\} \quad \textit{subject to} \quad Q(z') = 0 \ \textit{ if } \ \mathbf{1}_{\text{valid}}(z') = 0, \tag{13}$$

where $\beta > 0$. Then the unique maximizer is

$$Q^\star_\beta(z' \mid z) = \frac{K(z' \mid z)\, \mathbf{1}_{\text{valid}}(z')\, \exp\big(\beta S(z')\big)}{\sum_{y \in \mathcal{Z}} K(y \mid z)\, \mathbf{1}_{\text{valid}}(y)\, \exp\big(\beta S(y)\big)}. \tag{14}$$

Proof. Let $\mathcal{Z}_{\text{ok}} := \{z' \in \mathcal{Z} : \mathbf{1}_{\text{valid}}(z') = 1\}$ and restrict Q to be supported on $\mathcal{Z}_{\text{ok}}$. The objective in (13) can be written as

$$\sum_{z' \in \mathcal{Z}_{\text{ok}}} Q(z')\, S(z') - \frac{1}{\beta} \sum_{z' \in \mathcal{Z}_{\text{ok}}} Q(z') \log \frac{Q(z')}{K(z' \mid z)}. \tag{15}$$

Introduce a Lagrange multiplier λ enforcing $\sum_{z' \in \mathcal{Z}_{\text{ok}}} Q(z') = 1$. Differentiating with respect to $Q(z')$ and setting the derivative to zero yields

$$S(z') - \frac{1}{\beta}(\log Q(z') - \log K(z' \mid z) + 1) - \lambda = 0. \tag{16}$$

Rearranging gives

$$\log Q(z') = \log K(z' \mid z) + \beta S(z') + c, \tag{17}$$

for a constant c, implying

$$Q(z') \propto K(z' \mid z) \exp\big(\beta S(z')\big) \quad \text{on } \mathcal{Z}_{\text{ok}}. \tag{18}$$

Normalizing over $\mathcal{Z}_{\text{ok}}$ yields (14). The objective is strictly concave in Q, so the maximizer is unique. □

Corollary D.2 (Limits of guided kernels)**.** *Assume $K(\cdot \mid z)$ assigns positive mass to at least one valid state. Then: (i) as $\beta \to 0$, $Q^\star_\beta(\cdot \mid z)$ converges to $K(\cdot \mid z)$ restricted and renormalized to valid states; (ii) as $\beta \to \infty$, $Q^\star_\beta(\cdot \mid z)$ concentrates on valid maximizers of S within the support of $K(\cdot \mid z)$.*

Proof. Both statements follow directly from standard properties of the log-sum-exp normalization in (14). □

E PDC Model Architectures and Training Details

E.1 PepReDi Base Models

We adopt PepReDi [Chen et al., 2026c] as the discrete-flow base model for the PDC instantiation of CupOFMoCA, and train two variants that differ only in vocabulary and training corpus: **AA-PepReDi** (canonical 20 L-amino acids) and **UAA-PepReDi** (extended to D-amino acids and non-canonical residues). Both share architecture, optimizer, and rectification protocol; we describe the common pieces once and flag deltas where they apply.

Shared architecture and training. PepReDi is a discrete-flow model parameterized by a Diffusion Transformer applied in the MDLM regime [Peebles and Xie, 2023, Sahoo et al., 2024]. Per-token and per-timestep embeddings are produced by separate heads, summed, and propagated through stacked attention blocks with adaLN-modulated residual streams; a linear head emits per-position categorical logits. Both variants are trained with Adam ($\mathrm{lr} = 10^{-4}$, weight decay 10^{-5}; 2×10^{-5} for rectified runs), a 10-epoch warm-up, and cosine decay between 10^{-5} and the peak. Following ESM-2 [Lin et al., 2023], batches are formed by length-bucketing with a per-GPU token cap. We monitor conditional total correlation on a held-out subset (20 batches $\times$ 50 samples) and retain the checkpoint with the lowest value. At inference, both variants run 16 flow steps at temperature 1.0 to produce the prior μ_0 used downstream by CupOFMoCA.

AA-PepReDi. The canonical variant uses the ESM-2-650M tokenizer [Lin et al., 2023], yielding a vocabulary of 24 tokens (20 L-amino acids plus 4 special symbols), a model dimension of 512, 6 DiT layers, 8 attention heads, and a maximum sequence length of 100. Training data combines all peptides from PepNN [Abdin et al., 2022] and BioLip2 [Zhang et al., 2024] with 6-49-residue sequences from PPIRef [Bushuiev et al., 2024], partitioned 80:10:10 into train/validation/test. Optimization runs for 100 epochs on a single NVIDIA GPU (128 GB) under SLURM.

UAA-PepReDi. To support D-amino acids, head-to-tail (N-C) and disulfide (S-S) cyclization, and non-canonical residues required for realistic PDC carriers, we extend the tokenizer to 53 entries: the 4 specials and 20 L-residues from AA-PepReDi, 20 lowercase D-residue tokens, and 9 NCAAs available in our wet lab. Because no curated corpus combines all of these features, we synthesize training data with `p2smi` [Feller and Wilke, 2025a], sweeping residue composition over a 5×4 grid of NCAA fractions $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and D-residue fractions $\{0.1, 0.2, 0.3, 0.4\}$, with L-residues making up the remainder, yielding 20 composition combinations spanning the conjugate-relevant chemical space.The resulting corpus contains 150K sequences partitioned 80:10:10 into train/val/test. Architecture and optimization match AA-PepReDi except for the enlarged vocabulary.

PepReDi operates natively in amino-acid token space, but CupOFMoCA's objectives (binding affinity, conjugation validity) are defined on the assembled SMILES. Generated token sequences are therefore converted to peptide SMILES on the fly via three-letter codes [Feller and Wilke, 2025a], applying the appropriate cyclization closure when relevant, and the resulting molecule is passed to the scoring stack described in Section 3.4.

E.2 SMILES Base Model

We utilized the pre-trained SMILESReDi model from [Chen et al., 2026c] as our peptide SMILES generation model. SMILESReDi is a *rectified discrete flow* (ReDi) model [Yoo et al., 2025] built on top of discrete flow matching (DFM) Gat et al. [2024], using a Diffusion Transformer (DiT) architecture as its backbone. The model operates within a masked discrete generation framework and learns probability paths between source and target distributions rather than simulating a forward diffusion process.

SMILES sequences are tokenized using the PeptideCLM-23M tokenizer [Feller and Wilke, 2025b] and mapped to discrete tokens. Token embeddings and time-step embeddings are projected into continuous representations and processed through stacked DiT transformer blocks with Rotary Positional Embeddings (RoPE) and multi-head self-attention to capture long-range dependencies. The model outputs token-level transition probabilities via a final layer normalization followed by a linear projection.

E.2.1 Time-Dependent Bond-Aware Probability Path

A key feature of SMILESReDi is its bond-aware discrete probability path designed for peptide SMILES. Peptides contain a conserved backbone of alternating carbonyl and amide groups connected by peptide bonds, with variable side chains. To preserve this structure, SMILESReDi employs a time-dependent bond-aware probability path that transitions backbone bond tokens more slowly than side-chain tokens.

For each position j with bond indicator $b_j \in \{0, 1\}$, the time-t marginal of the probability path is defined as

$$p_t(x_t^{(j)} \mid x_0^{(j)}, x_1^{(j)}) = [b_j t^\gamma + (1-b_j)t]\delta_{x_1^{(j)}} + [1 - b_j t^\gamma - (1-b_j)t]\delta_{x_0^{(j)}}, \quad t \in [0,1],\ \gamma > 1, \tag{19}$$

ensuring that backbone tokens ($b_j = 1$) transition more slowly along the discrete flow path than non-bond tokens.

E.2.2 Training Configuration

The SMILESReDi model was trained and rectified following the configuration reported in [Chen et al., 2026c], and we use the released pre-trained weights without further fine-tuning. The configuration is detailed in Table E1.

Table E1: SMILESReDi training configuration.

Parameter	Value
Hardware	4×A6000 NVIDIA RTX 6000 Ada GPU (48 GB VRAM)
Training epochs	5
Optimizer	Adam
Learning rate	1×10^{-4}
Learning rate schedule	Cosine decay with 10% warmup
Minimum learning rate	1×10^{-5}
Model dimension	768
Transformer layers	8
Attention heads	8
Gradient clipping	1.0
Noising parameter γ	2.0

E.3 Binding Affinity Score Model

We employed the binding affinity prediction model from [Chen et al., 2026c] to score protein-peptide binding affinity. This model uses a reciprocal attention transformer architecture that utilizes unpooled token-level representations from the ESM-2 650M protein language model [Lin et al., 2023]. The architecture processes unpooled token embeddings from ESM-2 through convolutional layers followed by cross-attention layers, allowing the model to capture interaction patterns between protein and peptide sequences.

E.3.1 Model Architecture

The binding affinity predictor architecture is detailed in Table E2.

Table E2: Binding affinity model architecture.

Component	Configuration
Input	Unpooled token-level embeddings from ESM-2 650M
Convolutional layers	3 initial CNN kernel layers (dimension 384)
Cross-attention layers	4 layers (dimension 2048)
Prediction head	Shared prediction head (dimension 1024)

E.3.2 Training Details

As reported in the original work, the binding affinity data were split into a 0.8/0.2 ratio for training and validation, maintaining similar affinity score distributions across splits. The authors employed OPTUNA [Akiba et al., 2019] for hyperparameter optimization, tracking validation correlation scores as the objective metric. The training configuration is presented in Table E3.

Table E3: Binding affinity model training configuration.

Parameter	Value
Training epochs	50
Learning rate	3.84×10^{-5}
Dropout rate	0.15
CNN layers	3 initial layers (dimension 384)
Cross-attention layers	4 layers (dimension 2048)
Prediction head dimension	1024
Train/validation split	0.8/0.2
Validation performance	0.64 Spearman correlation

F PDC Conjugation Validity Constraints

During the generative workflow, a candidate PDC is first validated by the `PDCAnalyzer` for general structural feasibility, then subjected to conjugation-specific checks by the builder's validation methods, and finally scored for binding affinity and other therapeutic properties by downstream prediction models.

To ensure chemically valid PDCs from de-novo designed peptide components, we developed a specialized conjugation and validation pipeline using RDKit [Landrum et al., 2025]. Both the Aoa-Dau and SSCpt conjugation builders implement multi-layered validation comprising SMARTS-based pattern recognition, structural integrity checks, and valency analysis.

F.1 Aminooxyacetic acid - Daunorubicin Conjugates (Aoa-Dau)

F.1.1 Conjugation Reaction

Among the 21 PDC entries in PDCdb sharing the same peptide scaffold, Aminooxyacetic acid (Aoa) linker, and Daunorubicin (Dau) payload, four distinct conjugation strategies are represented, including direct conjugation via Lys side chain ε-amine, spacer-mediated conjugation, ranched conjugates and bis-conjugates at two Lys sites (additional Lys introduced via peptide modification).

We adopted the direct conjugation approach via Lys side chain NH_2, which is the most prevalent strategy among the database entries. Our implementation follows the architecture of PDC_00208 (GnRH-III-[8Lys(Dau=Aoa)]) as a reference template. To ensure site-selectivity, conjugation was restricted to the ε-amine of Lys side chains; the α-amine at the N-terminus was excluded as a potential reaction site.

Reaction Scheme. Conjugation proceeds via amide bond formation between the Lysine ε-amine (side chain $-NH_2$) on valid peptide, and the carboxyl group on Aoa (pre-conjugated to Dau via oxime bond):

$$\underbrace{\text{Peptide-Lys}}_{\text{targeting moiety}}\text{-NH}_2 \;+\; \text{HOOC-}\underbrace{\text{Aoa-Dau}}_{\text{linker-payload}} \xrightarrow{\text{amide coupling}} \text{Peptide-Lys-NH}{-}\text{CO-Aoa-Dau} \;+\; \text{H}_2\text{O} \tag{20}$$

Molecular Architecture. The complete PDC structure is represented schematically as:

$$\underbrace{\text{Peptide-Lys}}_{\text{De-novo Designed}}\text{-}\underbrace{\text{NH}{-}\text{C}(=\text{O})}_{\substack{\text{Amide}\\(\text{formed})}}\text{-}\underbrace{\text{CH}_2{-}\text{O}{-}\text{N}{=}\text{C}(\text{CH}_3)}_{\substack{\text{Aoa oxime}\\(\text{pre-formed})}}\text{-}\underbrace{\text{Dau}}_{\text{Payload}} \tag{21}$$

F.1.2 Attachment Site Compatibility Checks

The resulting PDC candidates need to pass all the checks in Table F1 for compatibility on the attachment site. For valency requirements, the amide nitrogen at the conjugation site must satisfy one of the following: a secondary amide, where N is bonded to 2 heavy atoms (one carbonyl C) and 1 H, or a tertiary amide, where N is bonded to 3 heavy atoms (one carbonyl C) and 0 H.

Method	SMARTS Pattern	Purpose
`has_aoa_amide_conjugation()`	`[NX3][CX3](=[OX1])[CH2][OX2][NX2]=[C]`	Confirms amide-CH_2-O-oxime linkage
`has_oxime_bond()`	`[C]-[O]-[N]=[C]`	Verifies Aoa-Dau linkage intact
`has_unreacted_aoa_carboxylic_acid()`	`[CX3](=[OX1])[OX2H1][CH2][OX2][NX2]=[C]`	Failure indicator: Free HOOC$-$CH$_2-$O$-$N$=$C
`find_amide_conjugation_sites()`	`[NX3:1][CX3:2](=[OX1])[CH2][OX2][NX2]=[C]`	Locates conjugation sites
`check_valency_at_conjugation_site()`	Structural analysis	Validates N valency

Table F1: AoaDau attachment site compatibility checks.

F.2 Disulfide-Camptothecin Conjugates (SSCpt)

F.2.1 Conjugation Reaction

Seven PDCdb entries share the same ERBB2-targeting peptide scaffold, disulfide linker, and camptothecin (CPT) payload. Their conjugation architectures include dimers with internal attachment and monomers with N-terminal or Lys side-chain attachment. Three linker-to-drug chemistries are observed: ester, carbonate, and carbamate.

We adopted the monomer architecture with Lys side chain NH_2 conjugation, following PDC_00294 (Z7, ester) and PDC_00295 (Z8, carbonate) as reference templates. To ensure site-selectivity, conjugation was restricted to the ε-amine of Lys side chains; the α-amine at the N-terminus was excluded as a potential reaction site.

Reaction Scheme. Conjugation proceeds via amide bond formation between the Lysine ε-amine (side chain $-NH_2$) on the valid peptide, and the carboxyl group on Cys (pre-conjugated to CPT via disulfide-linked spacer):

$$\underbrace{\text{Peptide-Lys}}_{\text{targeting moiety}}\text{-NH}_2 \;+\; \text{HOOC-}\underbrace{\text{Cys-SS-Spacer-CPT}}_{\text{linker-payload}} \xrightarrow{\text{amide coupling}} \text{Peptide-Lys-NH}{-}\text{CO-Cys-SS-Spacer-CPT} \;+\; \text{H}_2\text{O} \tag{22}$$

Molecular Architecture. The complete PDC structure is represented schematically as:

$$\underbrace{\text{Peptide-Lys}}_{\text{De-novo Designed}}\text{-}\underbrace{\text{NH}{-}\text{C}(=\text{O})}_{\substack{\text{Amide}\\(\text{formed})}}\text{-Cys-}\underbrace{\text{S}{-}\text{S}}_{\substack{\text{Disulfide}\\(\text{cleavable})}}\text{-Spacer-}\underbrace{\text{CPT}}_{\text{Payload}} \tag{23}$$

Two linker-to-drug chemistries are supported, differing in the bond connecting the spacer to Camptothecin:

- Ester (Z7): Cys$-$S$-$S$-$CH$_2-$CH$_2-$C(=O)$-$O$-$CPT
- Carbonate (Z8): Cys$-$S$-$S$-$CH$_2-$CH$_2-$O$-$C(=O)$-$O$-$CPT

F.2.2 Attachment Site Compatibility Checks

The resulting PDC candidates need to pass all the checks in Table F2 for compatibility on the attachment site. For valency requirements, the conjugation site must satisfy the following connectivity: Lys nitrogen with 3 bonds (NH connected to amide), carbonyl carbon with 3 bonds (C=O and two single bonds), and Cys α-carbon with 4 bonds (tetrahedral geometry).

Method	SMARTS Pattern	Purpose
`has_lys_cys_amide_conjugation()`	`[NX3][CX3](=[OX1])[CH]([CH2][SX2][SX2])[NX3]`	Confirms Lys-ε-N−C(=O)-C_α(Cys) pattern
`has_disulfide_bond()`	`[SX2]-[SX2]`	Verifies S-S linker intact
`has_unreacted_cys_carboxylic_acid()`	`[NH2][CH]([CH2][SX2][SX2])[CX3](=[OX1])[OX2H1]`	Failure indicator: Free Cys-COOH
`find_amide_conjugation_sites()`	`[NX3][CX3](=[OX1])[CH]([CH2][SX2][SX2])[NX3]`	Locates conjugation sites
`check_valency_at_conjugation_site()`	Structural analysis	Validates atom connectivity

Table F2: SSCpt attachment site compatibility checks.

F.3 Disulfanylpropanoic Acid - SN-38 Conjugates (DSPA-SN38)

F.3.1 Conjugation Reaction

Among the 9 PDC entries in PDCdb sharing the NRP1-targeting (RGD-motif) peptide scaffold family, 3-Disulfanylpropanoic Acid (DSPA, LIN00149) linker, and 7-Ethyl-10-hydroxycamptothecin (SN-38) payload, two distinct linker-to-drug attachment positions are represented: ester at the C-10 phenolic hydroxyl (most prevalent, e.g., PDC_02099) and ester at the C-20 tertiary aliphatic hydroxyl (e.g., PDC_02112). All entries adopt a monomer architecture employing a cleavable disulfide bond as the peptide-to-linker junction.

We adopted the monomer architecture with disulfide conjugation through the peptide cysteine thiol, following PDC_02099 (SN38(10)-CRGDK) and PDC_02112 (SN38(20)-CRGDK) as reference templates. To ensure site-selectivity, conjugation was restricted to free Cys side-chain thiol (−SH) groups; pre-existing disulfides and other sulfur-containing motifs were excluded as potential reaction sites. Both C-10 and C-20 ester variants are supported and pooled by default.

Reaction Scheme. Conjugation proceeds via disulfide bond formation between the cysteine thiol (side chain −SH) on the valid peptide, and the free terminal thiol on DSPA (pre-conjugated to SN-38 via an ester bond):

$$\underbrace{\text{Peptide-Cys}}_{\text{targeting moiety}}\text{-SH} \;+\; \text{HS-}\underbrace{\text{DSPA-SN38}}_{\text{linker-payload}} \xrightarrow{\text{thiol-disulfide exchange}} \text{Peptide-Cys-S}-\text{S-DSPA-SN38} \;+\; 2\,\text{H} \tag{24}$$

Molecular Architecture. The complete PDC structure is represented schematically as:

$$\underbrace{\text{Peptide-Cys}}_{\text{De-novo Designed}}\text{-}\underbrace{\text{S}-\text{S}}_{\substack{\text{Disulfide}\\\text{(formed, cleavable)}}}\text{-}\underbrace{\text{CH}_2-\text{CH}_2-\text{C}(=\text{O})-\text{O}}_{\substack{\text{DSPA ester}\\\text{(pre-formed)}}}\text{-}\underbrace{\text{SN-38}}_{\text{Payload}} \tag{25}$$

Two ester attachment positions on SN-38 are supported, differing in the hydroxyl that anchors the DSPA propanoate:

- C-10 (phenolic OH, default): $\text{Cys}-\text{S}-\text{S}-\text{CH}_2-\text{CH}_2-\text{C}(=\text{O})-\text{O}-\text{SN}_{38}(_{10})$
- C-20 (tertiary aliphatic OH): $\text{Cys}-\text{S}-\text{S}-\text{CH}_2-\text{CH}_2-\text{C}(=\text{O})-\text{O}-\text{SN}_{38}(_{20})$

F.3.2 Attachment Site Compatibility Checks

The resulting PDC candidates must pass all checks in Table F3 for compatibility on the attachment site. For valency requirements, both sulfur atoms in the newly formed disulfide must adopt an `SX2` configuration (each bonded to exactly one carbon and one sulfur, with zero residual hydrogens), the carbonyl carbon of the linker-drug ester must remain `CX3` (sp^2, bonded to an alkyl carbon, a doubly-bonded oxygen, and the ester oxygen), and no free thiol (−SH) may persist on either the peptide or the linker after coupling.

Method	SMARTS Pattern	Purpose
`has_disulfide_bond()`	`[SX2]-[SX2]`	Confirms Cys-S-S-linker bond formed
`has_ester_bond_to_drug()`	`[#6][CX3](=[OX1])[OX2][#6]`	Verifies DSPA-SN38 ester linkage intact
`has_sn38_core()`	`c1ccc2ncc3C(=O)OCc3c2c1`	Verifies SN-38 camptothecin scaffold preserved
`has_free_thiol()`	Atom-level scan (S, $\geq$1 H, degree 1)	Failure indicator: unreacted —SH remains
`find_conjugation_sites()` / `find_peptide_cysteines()`	Atom iteration (S, $\geq$1 H, degree 1)	Locates free Cys thiols on peptide
`check_attachment_site_compatibility()`	Structural analysis	Validates disulfide / no-free-thiol / ester jointly

Table F3: DSPA-SN38 attachment site compatibility checks.

F.4 PDC Analyzer

While the builder classes encapsulate linker-payload and conjugation reaction specific validation logic, the `PDCAnalyzer` remains conjugation-agnostic and can be applied to novel PDC designs under different conjugation reaction and / or linker-payload components

F.4.1 Hard Constraint Validation

The analyzer enforces two categories of hard constraints, both of which must be satisfied for a PDC to be considered structurally valid. Violation of any hard constraint results in immediate rejection of the candidate molecule.

Linker Feasibility Constraints. Linker regions in PDCs are particularly susceptible to structural infeasibility. The analyzer addresses two primary failure modes:

Ring-closure constraints are evaluated by identifying rings that are likely part of the linker rather than the peptide backbone or drug payload. Aromatic rings are excluded from linker classification as they typically belong to the drug core. The analyzer specifically searches for succinimide rings (five-membered rings containing nitrogen and two carbonyl groups), which commonly form from maleimide-based conjugation after thiol addition. For non-aromatic rings, size-based filtering is applied: three-membered rings are checked for sp^2 carbons participating in C=C double bonds (which introduce severe angle strain), four-membered rings are rejected if they contain more than two sp^2 centers, and rings larger than twelve atoms are flagged as synthetically implausible linker components.

Long carbon chain detection identifies excessively long aliphatic chains that would be synthetically challenging and prone to conformational heterogeneity. The analyzer uses a SMARTS pattern matching sixteen consecutive CH_2 groups, rejecting any PDC containing such chains.

Representational Validity Constraints. These constraints ensure that the generated SMILES string represents a chemically valid and processable molecule. Four checks are performed sequentially: (1) RDKit sanitization; (2) round-trip SMILES conversion, where the molecule is converted to SMILES and re-parsed to check if reconstruction could be conducted; (3) chemistry problem detection using RDKit's diagnostic tools, specifically checking for `AtomValenceException` and `KekulizeException` errors that indicate fundamental chemical inconsistencies; and (4) 2D coordinate generation, which serves as a final structural coherence test. Failure at any stage indicates that the molecule, while potentially representable as a string, does not correspond to a valid chemical structure.

F.4.2 Physicochemical Properties

Beyond binary validity assessment, the analyzer can compute eight molecular descriptors that are relevant for drug-likeness and pharmacokinetic profiling. These properties are toggled via boolean flags and include: molecular formula, topological polar surface area (TPSA), molecular complexity (BertzCT index), lipophilicity (XLogP via Wildman-Crippen method), heavy atom count, hydrogen bond donors and acceptors, and rotatable bond count.

G PDC Sampling Implementation

Table G1 defines the notation, and Tables G2 and G3 summarize the two sampling phases.

G.1 Notation

Table G1: Notation used in the algorithm description.

Symbol	Description
Φ_{θ^*}	Pretrained discrete flow model with parameters θ^*
$\mathcal{Z}$	Discrete token space (SMILES vocabulary)
μ_0	Prior distribution (uniform over $\mathcal{Z}^L$)
N	Number of flow steps for initial generation
K	Number of guided optimization steps
$S(z)$	Guidance score function
β	Guidance strength (temperature parameter)
$\mathbf{1}_{\text{valid}}(z)$	Validity indicator function
Z^*	Best sequence found during optimization
L	Sequence length

G.2 Two-Phase Generation Framework

Following Chen et al. [2026c], our approach separates generation into two phases, each serving a specific purpose. This design decouples *generation* (producing valid molecules) from *optimization* (improving target properties), allowing each phase to be tuned independently.

G.2.1 Phase 1: Initial Sequence Generation

The first phase rapidly generates an initial sequence Z_0 using unguided discrete flow matching. Starting from random noise, we iteratively denoise through the learned flow to obtain a molecular sequence satisfying the validity constraints. Running $N = 16$ steps provides sufficient denoising while remaining computationally efficient, and ensures the initial sequence lies on the data manifold. When the `require_valid_x0` flag is set, we regenerate Z_0 until the final sequence in this phase satisfies $\mathbf{1}_{\text{valid}}(Z_0) = 1$, enforcing that the optimization phase starts from a feasible point in the constraint manifold. This was adapted from the sampling process in Chen et al. [2026c], which was originally used to guarantee SMILES peptide validity.

Table G2: Phase 1 (initial generation) parameters.

Parameter	Value	Description
Flow Steps (N)	16	Number of denoising steps
Temperature	1.0	Softmax temperature for predictions
`require_valid_x0`	Enabled	Regenerate until $\mathbf{1}_{\text{valid}}(Z_0) = 1$
`max_x0_retries`	10	Maximum regeneration attempts

G.2.2 Phase 2: Guided Optimization

The second phase refines Z_0 through $K = 128$ steps of guided optimization using exponential tilting with an acceptance criterion, as in Algorithm 1. At each step, we select a random position, compute a tilted distribution over candidate tokens, and accept the proposal only if it improves (or maintains) the guidance score. This allows fine-grained control through local modifications, while the acceptance criterion prevents the optimization from degrading sequence quality.

Cascaded validity evaluation for residue-token representations. For AA- and UAA-tokenized generation, we use a compute-bounded cascade to avoid performing the expensive assembled-PDC validity check for every candidate proposal. At each guided update, let $\mathcal{C} = \mathrm{TopK}(p_{k+1})$ denote the

candidate set, with $|\mathcal{C}| \leq K$ and $K = 200$. We first decode each candidate into a peptide SMILES string and apply inexpensive checks for peptide validity and the presence of a compatible conjugation site. Binding affinity is evaluated for all candidates, after which the $\kappa = 8$ highest-affinity candidates among those passing the prefilter are retained. Full linker–payload conjugation and assembled-PDC validation are performed only for these retained candidates. Candidates that fail the prefilter or are not retained are assigned a zero validity indicator in the exponential tilt. SMILES tokenized generation does not use this cascade and applies the full scoring and validity procedure to all candidates in $\mathcal{C}$.

Log-Space Computation. For numerical stability, we compute the tilted distribution in log-space:

$$\tilde{p}(z') = \log p_k(z') + \beta \cdot S(z') + \log \mathbf{1}_{\text{valid}}(z') \tag{26}$$

$$\tilde{p}_{\text{stable}}(z') = \tilde{p}(z') - \max_{z''} \tilde{p}(z'') \tag{27}$$

$$p'(z') = \frac{\exp(\tilde{p}_{\text{stable}}(z'))}{\sum_{z''} \exp(\tilde{p}_{\text{stable}}(z''))} \tag{28}$$

Fallback Behavior. If exponential tilting produces invalid probabilities (due to all candidates being invalid or numerical instability), we fall back to sampling from the base distribution p_k.

Acceptance Criterion. After sampling $z^\dagger \sim p'$ and constructing the candidate sequence Z', we apply an acceptance criterion:

$$Z_{k+1} = \begin{cases} Z' & \text{if } S(Z') \geq S(Z_k) \\ Z_k & \text{otherwise} \end{cases} \tag{29}$$

This ensures the score never decreases, providing monotonic improvement guarantees. Combined with tracking Z^*, this guarantees we return the best sequence encountered during optimization.

Table G3: Phase 2 (guided optimization) parameters.

Parameter	Value	Description
Optimization Steps (K)	128	Number of guided refinement steps
Guidance Strength (β)	50.0	Exponential tilting temperature
Sequence Length (L)	80, 100	Generated sequence length (tokens)
Top-p	1.0	Nucleus sampling threshold
Top-K	200	Maximum candidates per position
Cascade size (κ)	8 (AA/UAA); not used (SMILES)	Candidates receiving full PDC validation
Position Selection	Random	Uniform random position each step
`use_acceptance`	Enabled	Reject proposals with $S(Z') < S(Z_k)$

Token Filtering. Invalid tokens are masked during both phases. The filtered indices $\{0, 1, 2, 3, 4, 585, 586\}$ correspond to special tokens: `[PAD]`, `[UNK]`, `[CLS]`, `[SEP]`, `[MASK]`, and reserved indices.

G.3 Computational Requirements

Sampling experiments were each conducted on a single A6000 NVIDIA GPU with 48GB of VRAM. For each sample, Phase 1 (16 flow steps) completes in under 1 minute; Phase 2 (128 optimization steps) takes 5-10 minutes depending on the sequence length and candidate size.

H PROTAC Base Model and Scorer Architectures

H.1 DeFoG Base Model

We use the pre-trained DeFoG discrete flow matching model [Qin et al., 2025] as our base PROTAC linker generator. DeFoG operates directly on molecular graphs rather than SMILES strings, representing nodes (atoms) and edges (bonds) as separate discrete random variables and learning a continuous-time discrete flow between a structured prior and the data distribution. The model uses a graph transformer architecture with separate node and edge update streams. We fine-tuned the released GuacaMol checkpoint of DeFoG for 100 epochs on $2{,}070$ PROTAC-DB linker graphs (258 validation, 260 test), adapting the prior to the PROTAC-linker distribution. The linker subgraph is the only region the flow modifies, with warhead and E3-ligand subgraphs held fixed throughout sampling (Section 3.3).

Table H1: Effect of PROTAC-specific fine-tuning of the DeFoG base model. Mean $\pm$ std over three seeds, 100 generations per seed.

System	Metric	Base	Fine-tuned
EGFR-VHL	Assembly validity (%)	78.7 ± 5.4	$\mathbf{85.0 \pm 3.6}$
	SA $\downarrow$	3.6 ± 0.2	$\mathbf{2.8 \pm 0.0}$
	MW in range (%)	48.7 ± 2.5	$\mathbf{94.7 \pm 1.9}$
BRD9-VHL	Assembly validity (%)	76.7 ± 3.3	$\mathbf{82.3 \pm 4.0}$
	SA $\downarrow$	3.6 ± 0.1	$\mathbf{2.8 \pm 0.0}$
	MW in range (%)	99.7 ± 0.5	100.0 ± 0.0
CDK6-CRBN	Assembly validity (%)	78.3 ± 5.3	$\mathbf{83.0 \pm 4.2}$
	SA $\downarrow$	3.6 ± 0.0	$\mathbf{2.8 \pm 0.0}$
	MW in range (%)	100.0 ± 0.0	100.0 ± 0.0

H.2 DegradeMaster Activity Predictor

For evaluation in Table 3 we use DegradeMaster [Liu et al., 2025], a neural classifier trained to predict whether a candidate PROTAC induces target degradation. DegradeMaster takes the assembled PROTAC SMILES as input and returns a probability in $[0, 1]$. We use the released pre-trained model and report mean predicted probability along with the fraction exceeding thresholds $0.3, 0.5, 0.7$. DegradeMaster is used solely for evaluation in this work; it is not used as a guidance objective during sampling.

We note that DegradeMaster and our fine-tuned generator both draw on PROTAC databases, so DegradeMaster constitutes an in-domain evaluation rather than a fully independent one. The memorization analysis (Table 4) rules out direct reproduction of training molecules, and the molecular-dynamics analysis (Section B.3) provides an evaluation that does not depend on the shared database.

I PROTAC Validity Constraints

We detail the implementation of the validity classifier $\mathbf{1}_{\text{valid}}$ and the soft connectivity scorer S_3 from Section 5.1. All checks are implemented in RDKit [Landrum et al., 2025].

I.1 Hard Validity Gate

A generated PROTAC is accepted as valid only if all three checks in Table I1 pass. Failure on any check returns $\mathbf{1}_{\text{valid}}(z) = 0$, which masks the candidate to zero probability under the tilted distribution during sampling.

I.2 Soft Connectivity Scorer

The connectivity scorer S_3 is computed even on states with $\mathbf{1}_{\text{valid}}(z) = 0$, providing a gradient toward single-component linkers during early sampling. Let $|\text{frag}|$ denote the number of connected

Check	Pattern / Procedure	Failure mode
Representational validity	`Chem.SanitizeMol(mol)` succeeds AND `MolFromSmiles(MolToSmiles(mol))` round-trips	Bad valences or aromaticity; molecule cannot be persisted
Liability filter	`[NX3][NX3]`, `[#7]=[#7]`, `[N+][O-]`, `[#7]-[OX2]`, `[NX3]C(=[OX1])[OX2H1]`	Hydrazide, azo, N-oxide, N-O bond, and carbamate patterns associated with instability, toxicity, or failure of amide assembly
Reactive termini count ≥ 2	Count of `[#6][CX3](=[OX1])[OX2H1]` (C-COOH) + `[#6][CX3H1](=[OX1])` (C-CHO) ≥ 2	Fewer than two reactive termini; warhead/E3 amide-coupling assembly impossible

Table I1: Hard validity gate checks. All three must pass for $\mathbf{1}_{\text{valid}}(z) = 1$.

fragments in the assembled graph, |largest| the size of the largest fragment, and |total| the total atom count. Then

$$S_3(z) = \begin{cases} 1.0 & \text{if } |\text{frag}| = 1, \\ \dfrac{|\text{largest}|}{|\text{total}| \cdot |\text{frag}|} & \text{otherwise.} \end{cases} \tag{30}$$

The score is halved if no path exists through the linker subgraph between the warhead anchor a^{wh} and the E3-ligand anchor a^{e3}.

I.3 Termini Conventions for Benchmark Systems

Each PROTAC system in our experiments specifies fixed warhead and E3-ligand anchor atoms at which amide-coupling assembly occurs. The conventions used are listed in Table I2.

System	Warhead anchor	E3 anchor
EGFR-VHL	Gefitinib-piperazine NH	VH032 NH_2 (tert-leucine)
BRD9-VHL	BI-7273-piperazine NH	VH032 NH_2 (tert-leucine)
CDK6-CRBN	Palbociclib-piperazine NH	Pomalidomide aryl-NH_2

Table I2: Fixed termini conventions for benchmark PROTAC systems.

J PROTAC Sampling Implementation

J.1 Algorithm

Algorithm 2 Exponential Tilting with Warmup and Best-So-Far Tracking

Input: trained discrete flow Φ_{θ^*}, prior μ_0 on $\mathcal{Z}$
time horizon T, total steps K, warmup steps K_{warm}
guidance score $S(z)$, strength β
Define time grid $0 = t_0 < t_1 < \cdots < t_K = T$
Sample $Z_0 \sim \mu_0$ s.t. $\mathbf{1}_{\text{valid}}(Z_0) = 1$; $Z^* \leftarrow Z_0$
(Warmup: unguided sampling from the pre-trained flow)
for $k = 0$ **to** $K_{\text{warm}} - 1$ **do**
Select a free linker site i
$p_{k+1} \leftarrow \Phi_{\theta^*}(Z_k)[\cdot, i]$
Sample $z^\dagger \sim p_{k+1}$ (no tilting, no acceptance)
$Z_{k+1} \leftarrow Z_k$ with site i set to $z^\dagger$
if $\mathbf{1}_{\text{valid}}(Z_{k+1}) = 1$ **and** $S(Z_{k+1}) > S(Z^*)$ **then**
$Z^* \leftarrow Z_{k+1}$ (track best-so-far during warmup)
end if
end for
(Guided phase: exponential tilting with hard validity mask)
for $k = K_{\text{warm}}$ **to** $K - 1$ **do**
Select a free linker site i (atom slot in $\mathcal{I}_{\text{link}}$ or bond entry incident to it)
$p_{k+1} \leftarrow \Phi_{\theta^*}(Z_k)[\cdot, i]$ (predicted distribution at site i)
for each $z' \in \text{TopK}(p_{k+1})$ **do**
$p'(z') \propto p_{k+1}(z') \exp(\beta\, S(z'))\, \mathbf{1}_{\text{valid}}(z')$
end for
Sample $z^\dagger \sim p'$
$Z' \leftarrow Z_k$ with site i set to $z^\dagger$ (atom-type or bond-type edit)
$Z_{k+1} \leftarrow Z'$ (no acceptance test; see Table J3)
if $\mathbf{1}_{\text{valid}}(Z_{k+1}) = 1$ **and** $S(Z_{k+1}) > S(Z^*)$ **then**
$Z^* \leftarrow Z_{k+1}$ (track best-so-far)
end if
end for
return Z^* (best PROTAC graph found)

J.2 Notation

Table J1: Notation used in the PROTAC sampling algorithm, with parameter values in Table J2.

Symbol	Description
Φ_{θ^*}	Pre-trained discrete flow model with parameters θ^*
$\mathcal{Z}$	Discrete state space (PROTAC molecular graphs)
μ_0	Prior distribution (empty linker, fixed warhead/E3)
K	Total number of guided sampling steps
K_{warm}	Number of unguided warmup steps
$S(z)$	Aggregate guidance score (Section 5.1)
β	Guidance strength (temperature parameter)
$\mathbf{1}_{\text{valid}}(z)$	Hard validity indicator
Z^*	Best PROTAC graph found during sampling

J.3 Two-Phase Sampling Framework

Sampling proceeds in two phases (Algorithm 2). The first phase performs unguided sampling from the pre-trained flow for K_{warm} steps, growing the linker outward from the anchors during the early-branching regime where the soft objectives provide limited discriminative signal (Section J.3.1). The second phase applies exponential tilting with hard validity masking for the remaining $K - K_{\text{warm}}$ steps, optimizing the aggregate objective $S(z)$ while maintaining chemical validity.

J.3.1 Phase 1: Warmup

The warmup phase samples directly from the pre-trained flow: $Z_{k+1} \sim \Phi_{\theta^*}(Z_k)$ for $k = 0, \ldots, K_{\text{warm}} - 1$. No tilting and no acceptance check are applied. The best-so-far state Z^* is tracked over warmup steps so that any high-scoring intermediate configurations are retained.

J.3.2 Phase 2: Guided Optimization

The guided phase applies exponential tilting with hard validity masking. At each step, a free linker site i (atom slot in $\mathcal{I}_{\text{link}}$ or bond entry incident to it) is selected, and the model's predicted distribution p_{k+1} at site i is tilted by the aggregate score S and masked by validity.

Table J2: Two-phase sampling parameters.

Parameter	Value	Description
Total steps (K)	1500	Total sampling steps (warmup + guided)
Warmup steps (K_{warm})	100	Unguided sampling from pre-trained flow
Guided steps	1400	$K - K_{\text{warm}}$, with tilting + no acceptance
Guidance strength (β)	5.0	Exponential tilting temperature
Top-K	50	Candidates per site for tilting
Temperature	1.0	Softmax temperature for predictions
Position selection	Random	Uniform random free linker site each step
`use_acceptance`	Disabled	Every tilted proposal accepted; best-so-far tracked

Log-Space Computation. For numerical stability, we compute the tilted distribution in log-space:

$$\tilde{p}(z') = \log p_k(z') + \beta \cdot S(z') + \log \mathbf{1}_{\text{valid}}(z'), \tag{31}$$

$$\tilde{p}_{\text{stable}}(z') = \tilde{p}(z') - \max_{z''} \tilde{p}(z''), \tag{32}$$

$$p'(z') = \frac{\exp(\tilde{p}_{\text{stable}}(z'))}{\sum_{z''} \exp(\tilde{p}_{\text{stable}}(z''))}. \tag{33}$$

Tilting and validity masking are restricted to the TopK candidates under p_{k+1} to limit per-step scorer evaluation cost.

Fallback Behavior. If exponential tilting produces invalid probabilities (all candidates fail validity, or numerical underflow), we fall back to sampling from the base distribution p_{k+1} restricted to valid candidates. If no valid candidate exists in the TopK, the step is rejected and $Z_{k+1} = Z_k$.

Acceptance Criterion (disabled). We tested the monotonic acceptance criterion in the PROTAC setting. Requiring $S(Z') \geq S(Z_k)$ reduced assembly validity by 24-26 percentage points and slightly increased SA (Table J3). The final sampler accepts each tilted proposal and returns the highest-scoring valid state encountered. The PDC sampler (Appendix C) retains the acceptance criterion.

Table J3: Monotonic acceptance ablation. Assembly validity (%), mean $\pm$ std over three seeds, 100 linkers per seed.

System	Acceptance on	Acceptance off
EGFR-VHL	58.7 ± 3.4	$\mathbf{85.0 \pm 3.6}$
BRD9-VHL	58.3 ± 4.1	$\mathbf{82.3 \pm 4.0}$
CDK6-CRBN	58.7 ± 2.9	$\mathbf{83.0 \pm 4.2}$

J.4 Objective Weights

The aggregate objective $S(z) = \sum_{j=1}^{3} w_j S_j(z)$ uses the weights in Table J4, selected by the ablation in Appendix J.5. The connectivity weight $w_3 = 4.0$ is elevated so that the topological gradient

toward connected, anchor-bridging linkers dominates early in sampling, where partial assemblies score similarly under S_1-S_2.

Table J4: Soft objective weights for the aggregate score $S(z)$.

Scorer	Description	Weight
S_1	Synthetic accessibility ramp	4.0
S_2	PROTAC-DB chemotype closeness	2.0
S_3	Soft connectivity gradient	4.0

J.5 Objective Weight Ablation

We swept each objective weight independently on EGFR-VHL, generating 100 linkers across three seeds $(42/43/44)$ per setting while holding the remaining weights at 1. Tables J5, J6, and J7 report the completed sweep. Every row shares the weight-1 reference configuration, and the term below the rule in each table is the one excluded from the final objective. A PROTAC-likeness classifier term is omitted from all three: it acts as a $0/1$ gate, so its weight cannot change candidate ranking and every metric is identical to the reference at each weight.

We selected the per-term argmax of this sweep for the final configuration: connectivity 4, SA 4, chemotype 2, and fingerprint 0. Combining the three retained terms reached $85.0 \pm 3.6\%$ assembly validity (Table 3), above any single-term setting in the sweep.

Table J5: Objective weight sweep on EGFR-VHL: assembly validity (%), higher is better. Seeds $42/43/44$, 100 linkers per setting, mean $\pm$ std over seeds. **Bold**: value used in the final configuration.

Term	0	1	2	4
Connectivity	36.7 ± 4.7	60.7 ± 8.5	72.0 ± 5.3	$\mathbf{75.3 \pm 5.5}$
Chemotype closeness	51.0 ± 7.9	60.7 ± 8.5	$\mathbf{64.7 \pm 10.0}$	64.3 ± 8.0
Synthetic accessibility	62.7 ± 5.7	60.7 ± 8.5	60.7 ± 9.5	$\mathbf{55.0 \pm 8.9}$
Morgan fingerprint	$\mathbf{64.3 \pm 8.0}$	60.7 ± 8.5	52.3 ± 9.1	50.7 ± 5.5

Table J6: Objective weight sweep on EGFR-VHL: SA score, lower is better. Same runs as Table J5.

Term	0	1	2	4
Connectivity	2.94 ± 0.03	3.03 ± 0.11	3.07 ± 0.20	$\mathbf{3.11 \pm 0.07}$
Chemotype closeness	2.96 ± 0.11	3.03 ± 0.11	$\mathbf{3.11 \pm 0.14}$	3.21 ± 0.16
Synthetic accessibility	3.36 ± 0.06	3.03 ± 0.11	2.80 ± 0.12	$\mathbf{2.64 \pm 0.08}$
Morgan fingerprint	$\mathbf{3.00 \pm 0.07}$	3.03 ± 0.11	3.06 ± 0.12	3.24 ± 0.10

Table J7: Objective weight sweep on EGFR-VHL: mean DegradeMaster probability over valid generations, higher is better. Same runs as Table J5.

Term	0	1	2	4
Connectivity	0.518 ± 0.022	0.481 ± 0.022	0.498 ± 0.029	$\mathbf{0.496 \pm 0.031}$
Chemotype closeness	0.522 ± 0.021	0.481 ± 0.022	$\mathbf{0.489 \pm 0.046}$	0.422 ± 0.046
Synthetic accessibility	0.469 ± 0.033	0.481 ± 0.022	0.533 ± 0.026	$\mathbf{0.536 \pm 0.028}$
Morgan fingerprint	$\mathbf{0.506 \pm 0.053}$	0.481 ± 0.022	0.470 ± 0.057	0.447 ± 0.012

Removing connectivity guidance reduced assembly validity to 36.7%; weight 4 reached 75.3%. Chemotype closeness produced a smaller gain that saturated at weight 2. Increasing the synthetic-accessibility weight from 0 to 4 reduced SA from 3.36 to 2.64, reduced assembly validity by roughly 8 points, and increased DegradeMaster from 0.469 to 0.536. We retained this setting because synthetic accessibility is relevant to experimental follow-up.

We excluded the Morgan fingerprint term. Across the four weights, assembly validity fell monotonically from 64.3% to 50.7%, SA increased from 3.00 to 3.24, and DegradeMaster decreased from 0.506 to 0.447. Removal gave the best value for all three metrics. The individual DegradeMaster cells overlap within one standard deviation, so we interpreted the result as a monotone trend across four weights and three metrics. This conclusion applies to the EGFR-VHL sweep. On BRD9-VHL and CDK6-CRBN the same four settings are flat (86.0-86.2% assembly validity, DM 0.530-0.535). We attribute the effect to a pull toward training-set chemotypes, which conflicts with the novelty results in Table 4; across all settings we observed 0% exact memorization of training molecules.

We also ablated hard constraints and soft guidance (Table J8). Soft guidance alone reached $0.3 \pm 0.5\%$ assembly validity, hard constraints alone reached $11.0 \pm 3.7\%$, and their combination reached $85.0 \pm 3.6\%$. The combined result supports the joint use of feasibility constraints and directional guidance.

Table J8: Soft guidance and hard constraint ablation on EGFR-VHL. Assembly validity (%), mean $\pm$ std over three seeds, 100 linkers per seed. The full configuration reaches 82.3-85.0% across all three benchmark systems (Table 3).

Configuration	**Assembly validity (%)**
Soft guidance only	0.3 ± 0.5
Hard constraints only	11.0 ± 3.7
Both	**85.0 $\pm$ 3.6**

J.6 Computational Requirements

Sampling experiments were conducted on a single NVIDIA A6000 GPU with 48 GB of VRAM. We generated 100 PROTACs per target in one parallel batch. The reported sampling times cover each complete batch.

Under a matched compute budget, CupOFMoCA generated 0.247 assembly-valid PROTACs per second, compared with 0.181 for DiffPROTACs, 0.072 for LinkerNet, and 0.000 for DiffLinker. We generated each batch of 100 linkers on one NVIDIA RTX A6000 with an AMD EPYC 7662 CPU and eight CPU threads. CupOFMoCA completed the batch in 356 s, compared with 442, 636, and 432 s for DiffPROTACs, LinkerNet, and DiffLinker, respectively.

CPU-side constraint evaluation limited throughput. Each guided step evaluated the classifier and soft-scorer stack (RDKit sanitization, SMARTS pattern matching, SA score, MW/logP/RotBonds windows, connectivity check) on every candidate type; GPU forward passes consumed under 20% of per-step wall-clock time. Vectorizing the CPU-bound checks reduced the original 8-11 minute implementation to 356 seconds with byte-identical outputs and unchanged validity and quality metrics.

K Impact Statement

This work introduces a generative framework for designing molecular conjugates such as peptide-drug conjugates (PDCs) and PROTACs, which are widely used in targeted therapeutics. By improving the ability to generate chemically valid and functionally consistent conjugates, this approach may accelerate early-stage drug discovery and reduce experimental screening costs. Potential benefits include more efficient design of targeted therapies for cancer and other diseases. However, as with all generative methods in molecular design, there is a risk of unintended misuse, such as the generation of harmful or toxic compounds. In this work, we restrict generation through chemical validity constraints and domain-specific objectives, and do not release any experimentally validated compounds. Future work should consider incorporating additional safety filters and oversight when deploying such models in practical settings.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction (Sec. 3) state the paper's three contributions: (i) a coupled discrete flow framework for molecular conjugate assembly that enforces cross-component compatibility during generation, (ii) an objective-guided sampling scheme under hard validity constraints, and (iii) a unified instantiation across PDC and PROTAC modalities. Each claim is supported in the paper: the coupled flow formulation in Sec. 3 (Sec. 3.1–3.3); the validity-gated exponential-tilting sampler in Sec. 3.6, Appendix C, and Algorithm 2; PDC results across three targets in Sec. 4 (Tables 1, 2); and PROTAC results in Sec. 5.2 (Tables 3, 4).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in the Discussion section. The molecular-dynamics analysis addresses linker conformational dynamics and ternary-geometry compatibility for four generated EGFR–VHL candidates, but does not capture full ternary-complex formation or cellular context, and DegradeMaster is an in-domain post-hoc predictor rather than an experimental readout; experimental validation will be required to assess stability, target engagement, and functional efficacy of generated conjugates. PROTAC baselines are evaluated as single batches while CupOFMoCA is run across three seeds, so cross-method differences carry uneven uncertainty. The Discussion also flags the current assumption of fixed linker–payload chemistry as a limitation to be relaxed in future work, and identifies extension to additional conjugation modalities and pharmacokinetic objectives as open directions. Computational requirements per sample are reported in Appendix G.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results are stated formally with their assumptions and accompanied by complete proofs in Appendix D. Specifically, Theorem D.1 (exponential tilting solves KL-regularized objective improvement under hard validity constraints) is stated with explicit assumptions on the base kernel $K(\cdot \mid z)$, the score S, and the validity indicator $\mathbf{1}_{\text{valid}}$, and is proved via Lagrangian optimization with a strict-concavity argument that establishes uniqueness. Corollary D.2 characterizes the $\beta \to 0$ and $\beta \to \infty$ limits of the guided kernel and is proved from log-sum-exp normalization properties.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides full details required for reproduction. Base-model architectures, optimizers, schedules, and training corpora for AA-PepReDi, UAA-PepReDi, and SMILESReDi are described in Appendix E.1 (with the SMILESReDi configuration in Table E1); the binding affinity oracle is documented with architecture (Table E2) and training configuration (Table E3). Sampling is fully specified by Algorithm 2 together with Phase 1 and Phase 2 hyperparameters in Tables G2 and G3. Hard validity constraints used during generation are formalized in Sec. 3.4 for PROTACs and instantiated per conjugation chemistry in Appendix F (Tables F1, F2, F3). Benchmark targets, linker–payload pairings, and candidate counts (100 generations per seed across three seeds for CupOFMoCA and single batches of $N = 100$ for baselines on PROTACs; top-20 over 100 generations for PDCs) are specified in Sec. 4 and Sec. 5.2. An anonymous code repository is provided as part of the submission.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymous code repository is provided alongside the submission, containing the CupOFMoCA framework, the per-conjugation builders and `PDCAnalyzer` for PDCs, the PROTAC validity classifier, the separate PROTAC-likeness gate, the three soft scorers S_1–S_3 defined in Sec. 3.4, and the two-phase sampling pipeline of Algorithm 2. Experiments use publicly available resources, including PDCdb [Sun et al., 2025], PROTAC-DB, the EGFR–VHL/BRD9–VHL/CDK6–CRBN PROTAC systems, PepNN [Abdin et al., 2022], BioLip2 [Zhang et al., 2024], and PPIRef [Bushuiev et al., 2024], all cited in the corresponding sections. Pre-trained components (PepReDi/SMILESReDi base flows from Chen et al. [2026c], the ESM-2-650M tokenizer/encoder [Lin et al., 2023], the binding-affinity oracle, and DegradeMaster [Liu et al., 2025]) and baselines (LinkerNet [Guan et al., 2023], DiffPROTACs [Li et al., 2024], DiffLinker [Igashov et al., 2024], PepTune [Tang et al., 2025a], AReUReDi [Chen et al., 2026c]) are publicly available and cited.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Full training and test details are reported across the appendices. Base-model training settings (Adam with $\mathrm{lr} = 10^{-4}$, weight decay 10^{-5}, 2×10^{-5} for rectified runs, 10-epoch warmup, cosine decay, 16-step inference, and 80:10:10 train/validation/test splits for the AA/UAA corpora) are given in Appendix E.1; SMILESReDi configuration is in Table E1; the binding affinity oracle uses an Optuna-tuned configuration described in Tables E2 and E3. Sampling-time settings for PDCs (Phase 1: $N = 16$ flow steps at temperature 1.0 with `require_valid_x0`; Phase 2: $K = 128$ guided steps, $\beta = 50.0$, top-$K = 200$, random position selection, acceptance criterion enabled) are reported in Tables G2 and G3. PROTAC sampling uses 1500 total steps (100 unguided warmup, 1400 guided), $\beta = 5.0$, top-$K = 50$, random position selection, and monotonic acceptance disabled, because requiring non-decreasing scores reduced assembly validity (Table J2; Table J3). PROTAC scorers, gating ranges, and the weights $(w_1, w_2, w_3) = (4.0, 2.0, 4.0)$ selected by the sweep in Appendix J.5 are documented in Sec. 3.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Quantitative results report mean and standard deviation across multiple generated samples per condition. PDC tables (Tables 1, 2, A4) report top-20 binding affinity as mean (std) over 100 generation runs per method per target. Full-valid-set tables (Tables A5, A6, A7, and per-target tables A10–A18) extend this to all valid conjugates. PROTAC results (Table 3) report validity as mean $\pm$ std across three seeds of 100 generations for CupOFMoCA and single batches of $N = 100$ for baselines, with DegradeMaster as mean (std) over assembly-valid generations together with threshold-crossing fractions; Table 4 reports diversity and training-set similarity over the same pooled sets. Figure A1 additionally visualizes binding-affinity error bars (± 1 std) for the joint validity–binding tradeoff, and Figure A3 reports physicochemical property distributions as box plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in Appendix G.3 and the model-training appendices. AA-PepReDi training (Appendix E.1) uses a single NVIDIA GPU with 128 GB of GPU memory under SLURM. SMILESReDi training (Table E1) uses a $4\times$NVIDIA RTX A6000 Ada system with 48 GB of VRAM per GPU. CupOFMoCA sampling (Appendix G.3) is performed on a single NVIDIA A6000 GPU with 48 GB of VRAM, with Phase 1 (16 flow steps) completing in under one minute and Phase 2 (128 guided optimization steps) taking 5–10 minutes per sample depending on sequence length and candidate size. Matched-hardware PROTAC throughput is reported in Appendix J.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. The work does not involve human subjects or sensitive personal data, all datasets and pre-trained models used are publicly available and properly cited, and the submission preserves anonymity through the anonymous code repository. The Impact Statement explicitly addresses dual-use considerations for accelerated therapeutic conjugate design.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper includes a dedicated Impact Statement in the Appendix that discusses both the positive societal impacts of CupOFMoCA, including its ability to generate chemically valid molecular conjugates that preserve predicted activity at the assembled-system level and to extend across PDC and PROTAC modalities for therapeutic discovery, and the negative considerations, including that improved *in silico* validity and predicted activity do not guarantee *in vivo* safety, target engagement, or efficacy, and that accelerated conjugate design can raise safety and dual-use considerations. The Impact Statement explicitly notes that CupOFMoCA must be coupled with careful experimental validation, safety screening, and responsible deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new foundation models, datasets, or general-purpose generators that pose a high risk for misuse. Released artifacts consist of the CupOFMoCA sampling framework, conjugation builders, and validity analyzers, together with domain-specific models derived from existing public resources. The PDC part uses the publicly available base models (PepReDi/SMILESReDi from Chen et al. [2026c]) and publicly available conjugate database PDCdb. The PROTAC generator is fine-tuned from publicly released GuacaMol checkpoint of DeFoG [Qin et al., 2025] using public database PROTAC-DB. The Impact Statement notes that downstream therapeutic use requires careful experimental validation and safety screening before deployment.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets used in the paper are properly credited via citations to the original works, including PDCdb [Sun et al., 2025], PROTAC-DB, PepNN [Abdin et al., 2022], BioLip2 [Zhang et al., 2024], PPIRef [Bushuiev et al., 2024], ESM-2-650M [Lin et al., 2023], the PepReDi and SMILESReDi base flows and PeptideCLM-23M tokenizer [Chen et al., 2026c], the p2smi toolkit [Feller and Wilke, 2025a], RDKit [Landrum et al., 2025], Optuna [Akiba et al., 2019], DegradeMaster [Liu et al., 2025], and the LinkerNet [Guan et al., 2023], DiffPROTACs [Li et al., 2024], DiffLinker [Igashov et al., 2024], PepTune [Tang et al., 2025a], and AReUReDi [Chen et al., 2026c] baselines.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced are the CupOFMoCA framework and the associated conjugation builders, validity analyzers, and PROTAC scorers. The framework is documented in Sec. 3, Sec. 3.4, and Algorithm 2; the per-conjugation builders and `PDCAnalyzer` for Aoa–Dau, SSCpt, and DSPA–SN-38 are documented in Appendix F, including SMARTS patterns and valency requirements (Tables F1, F2, F3); the PROTAC validity classifier, the separate PROTAC-likeness gate, and the three soft scorers S_1–S_3 are documented in Sec. 3.4; and the sampling pipeline is documented in Appendix C and Tables G2 and G3. The accompanying anonymous code repository contains corresponding implementations.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments are computational and use publicly available molecular and peptide data resources.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methodology. The framework concerns coupled discrete flows and validity-gated exponential-tilting sampling for molecular conjugate assembly, with Diffusion Transformer backbones for the PepReDi and SMILESReDi base flows. The pre-trained ESM-2-650M sequence encoder is used only as a tokenizer and embedding source for the binding-affinity oracle, as cited in Appendix E.1.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.