NAVIGATING MULTIOBJECTIVE DRUG DISCOVERY TRADE-OFFS WITH PREFERENCE-GATED GFLOWNETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-objective molecular generation faces a double challenge: finding molecules that survive demanding screens and changing what is sampled when a scientist changes priorities. A broad Pareto frontier alone does not establish such control. We introduce PREFGATEGFN, a GFlowNet that uses preference-dependent feature gates to modulate molecular construction-action scores. Gates can change action rankings, while optional learned temperatures adjust sampling concentration; dynamic reward normalization addresses coverage as a separate training choice. We examine these roles using exact-grid distributions, matched molecular controls, and preference interventions on frozen policies. In a three-objective fragment task, the full system achieves higher mean top-10 reward, diversity, and hypervolume than the evaluated MOGFN and HN-GFN systems. A three-seed ablation raises mean hypervolume from 0.668 with normalization and curriculum disabled to 0.885 with normalization alone. In a comparison of complete systems with different training recipes, PREFGATEGFN achieves substantial yield under a demanding fourteen-objective screen across five protein targets: an average of 133 distinct molecules per 1,000 draws satisfying docking, physicochemical, and twelve predicted ADMET criteria, outperforming MOGFN by factors of 1.93 in qualifying molecules and 2.10 in qualifying scaffolds. Together with its measured preference response, these results support steering molecular generation toward diverse, computationally screened candidates for experimental follow-up.

1 INTRODUCTION

Drug candidates are selected on profiles, not isolated scores (Segall, 2012; 2014). A useful molecule must combine target affinity with suitable physicochemical properties, permeability, and an acceptable safety profile, while remaining synthetically accessible. These requirements can be weakly correlated or conflicting, making molecular design a multi-objective optimization problem (Segall, 2012; 2014; Nicolaou & Brown, 2013). Molecular generative models therefore need to consider these objectives jointly while allowing users to specify their relative importance (Jin et al., 2020). This creates two complementary requirements: the generator must discover a broad range of useful property profiles, and it must let a user steer sampling toward the profiles relevant to a particular discovery decision.

Generative flow networks (GFlowNets) provide a foundation for this goal by learning stochastic policies that construct candidates sequentially (Bengio et al., 2021; 2023). Given a condition c and a non-negative reward $R(x \mid c)$ with a finite, positive normalizing constant, a conditional GFlowNet targets a distribution over completed candidates x proportional to that reward, $p(x \mid c) \propto R(x \mid c)$. Multi-objective GFlowNets (MOGFNs) extend this reward-based framework to generation across different objective trade-offs (Jain et al., 2023). In their preference-conditioned formulation, a preference vector specifies how objective rewards are combined into a scalar reward according to their relative importance. Each preference therefore defines a target distribution over candidates, and a single policy learns this family of distributions by conditioning on the preference vector.

However, defining a preference-conditioned target does not ensure that a finite trained policy responds effectively to changes in preference. Common implementations provide preferences as an additional input to a shared policy network, which must learn how those preferences should influence construction decisions. When objectives differ in their scales or optimization difficulty, high aggregate reward can coexist with weak changes in the properties that a user prioritizes. This motivates a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

conditioning architecture that connects user preferences more directly to the evaluation of candidate actions, together with direct tests of whether changing those preferences changes generated outcomes.

We introduce PREFGATEGFN, a preference-gating architecture for molecular design within the preference-conditioned multi-objective GFlowNet framework. Our key idea is to let preferences determine how learned features contribute to the evaluation of candidate construction steps. A controller maps the preference vector to multiplicative feature gates that rescale the representations used for action scoring; in a bilinear action head, gates can modulate both state and action embeddings. The controller and policy are optimized within the GFlowNet objective to approximate the reward-proportional target distribution for each preference. An optional preference-dependent temperature separately controls action probabilities. Feature gates can change action rankings, whereas a positive scalar temperature preserves rankings within an action head at a fixed state. The gating mechanism provides a direct preference pathway without itself changing the scalarized reward or requiring a separate generator for each trade-off. We combine this interface with dynamic reward normalization in the fragment setting and with synthesis-aware construction in RxnFlow (Seo et al., 2025).

We evaluate PREFGATEGFN on controlled grids and in fragment-based and synthesis-aware molecular generation. The objective-count analysis in Appendix A shows lower mean distribution-matching error than MOGFN on the four- through eight-objective tasks, with performance close to HN-GFN. On the fragment-based benchmark (Bengio et al., 2021; Jain et al., 2023), PREFGATEGFN achieves higher mean top-10 reward, top-10 diversity, and hypervolume than the evaluated MOGFN and HN-GFN systems. Direct preference sweeps show larger SEH and QED responses than MOGFN on the tested edges. These findings distinguish responsiveness from candidate quality: PREFGATEGFN combines useful preference control with stronger aggregate quality and objective-space coverage, rather than dominating every response metric.

When integrated into RxnFlow (Seo et al., 2025), PREFGATEGFN generates many distinct molecules satisfying joint computational requirements across five protein targets. On the fourteen-objective panel that includes docking, a physicochemical criterion, and twelve predicted ADMET endpoints it produces $1.93\times$ as many distinct qualifying molecules and $2.10\times$ as many qualifying scaffolds as MOGFN-PC on average at the same evaluation draw budget. Appendix J extends the comparison to nine-objective training streams from FragGFN-SA (Bengio et al., 2021), RxnFlow (Seo et al., 2025), and SynFlowNet (Cretu et al., 2025): PREFGATEGFN yields more distinct all-seven-ADMET and fully qualifying candidates over the same recorded iteration window. Full-system training recipes differ, so these results establish practical system-level performance rather than isolating gating alone.

Our contributions are summarized as follows:

- We introduce PREFGATEGFN, a preference-conditioned GFlowNet architecture that uses learned multiplicative gates to modulate the features used for action scoring, with a separate optional temperature.
- We analyze preference-conditioned distributions and molecular responses through explicit preference interventions, distinguishing responsiveness from aggregate candidate quality and examining conditioning components under matched controls. A full-budget training ablation separates normalization from curriculum and identifies normalization as the main source of the observed coverage improvement within the tested gated architecture.
- We demonstrate improved reward and objective-space coverage in fragment-based generation and increased yields of distinct molecules satisfying joint docking, physicochemical, and predicted ADMET requirements in synthesis-aware design, including a fourteen-objective panel across five protein targets.

2 PRELIMINARIES

2.1 GFLOWNETS

A generative flow network (GFlowNet) learns a stochastic policy that constructs objects sequentially and samples completed objects in proportion to a specified reward (Bengio et al., 2021; 2023). The construction process is represented by a directed acyclic graph, whose states are partial objects and whose edges are valid construction actions. Starting from an initial state s_0, a forward policy $P_{F,\theta}$

generates a trajectory $\tau = (s_0 \to s_1 \to \cdots \to s_T = x)$ ending at an object $x \in \mathcal{X}$, where multiple trajectories may lead to the same terminal state.

Given a condition c and a positive reward function $R(x \mid c)$, the target distribution is

$$p^\star(x \mid c) = \frac{R(x \mid c)^\beta}{Z(c)}, \qquad Z(c) = \sum_{x' \in \mathcal{X}} R(x' \mid c)^\beta. \tag{1}$$

Here $\beta > 0$ is a fixed reward exponent and $Z(c)$ is assumed finite and positive. The terminal probability of an object is the sum of the probabilities of all generative trajectories leading to it.

2.2 MULTI-OBJECTIVE GFLOWNETS

Multi-objective GFlowNets (MOGFNs) extend reward-proportional generation to tasks in which candidates are evaluated against multiple, potentially conflicting objectives (Jain et al., 2023). For m objectives, let $\mathbf{r}(x) = (r_1(x), \ldots, r_m(x))$ collect their rewards, with larger values indicating more desirable outcomes. Logarithmic scalarizations use a positive floor when an objective can be zero. A scalarization function combines these objective rewards into a single positive reward that defines the target sampling distribution. To accommodate different trade-offs, the policy is conditioned on a non-negative preference vector $\mathbf{w} = (w_1, \ldots, w_m)$, with $\sum_{i=1}^m w_i = 1$, that specifies the relative importance of the objectives. The preference vector $\mathbf{w}$ determines the scalarized reward $R(x \mid \mathbf{w})$ and serves as the condition c in Equation 1. A single policy is trained across sampled preferences, keeping $\mathbf{w}$ fixed along each trajectory and conditioning both the forward policy and the learned normalizer on it.

Reward scalarization. The scalarization determines the family of reward functions learned across preferences and can be chosen independently of the conditioning architecture. Two choices are weighted-sum and weighted-log-sum scalarization. Weighted-sum scalarization combines the objective rewards additively:

$$R(x \mid \mathbf{w}) = \sum_{i=1}^m w_i r_i(x). \tag{2}$$

Weighted-log-sum scalarization combines their logarithms, giving a weighted geometric reward:

$$R(x \mid \mathbf{w}) = \prod_{i=1}^m r_i(x)^{w_i}, \qquad \log R(x \mid \mathbf{w}) = \sum_{i=1}^m w_i \log r_i(x). \tag{3}$$

The grid and fragment tasks use the weighted sum. The finite reaction-library control uses the geometric reward with a 10^{-4} component floor; the synthesis-aware protocols are specified in Appendix F. The geometric reward penalizes a low component with positive weight, but it is not equivalent to the hard conjunction used for reporting screening yield. The interpretation of a preference vector therefore depends on both the objective reward transformations and the chosen scalarization.

The scalarized reward defines the target distribution, while the conditioning architecture determines how preferences influence construction decisions. Let $\mathbf{h}_\theta(s)$ denote the state representation and f a prediction head that outputs action logits. The softmax normalizes these logits over the valid construction actions at state s.

Concatenation conditioning. A common conditioning approach concatenates a preference encoding $\mathbf{e}(\mathbf{w})$ with the state representation:

$$P_F(\cdot \mid s, \mathbf{w}; \theta) = \mathrm{softmax}(f_\theta(\mathrm{concat}(\mathbf{h}_\theta(s), \mathbf{e}(\mathbf{w})))), \tag{4}$$

where θ collectively denotes the parameters of the state encoder and prediction head, which are shared across preferences.

Hypernetwork-based conditioning. Hypernetwork-based GFlowNets (HN-GFN) generate the prediction head parameters from the preference vector using a hypernetwork $\phi(\cdot)$ (Zhu et al., 2023):

$$P_F(\cdot \mid s, \mathbf{w}; \theta, \phi) = \mathrm{softmax}\big(f_{\phi(\mathbf{w})}(\mathbf{h}_\theta(s))\big), \tag{5}$$

where $\phi(\mathbf{w})$ supplies the preference-dependent parameters of the prediction head, while the state encoder parameters θ are shared across preferences.

Training objective. With backward policy P_B, trajectory balance uses the residual

$$\delta(\tau, \mathbf{w}) = \log Z_\theta(\mathbf{w}) + \sum_t \log P_{F,\theta}(a_t \mid s_t, \mathbf{w}) - \beta \log R(x \mid \mathbf{w}) - \sum_t \log P_B(s_t \mid s_{t+1}, \mathbf{w}). \quad (6)$$

For a fixed reward, its zero-residual solution satisfies Equation 1 (Malkin et al., 2022). The fragment study uses subtrajectory balance (Madan et al., 2023) with a conditional state-flow estimator; the finite-library control minimizes the squared residual. Reaction runs use the stated squared or Huber regression variant. Training settings and differences between complete systems are reported with each protocol.

3 PREFERENCE-GATED GFLOWNETS

PREFGATEGFN connects the objective preference directly to the features used to score candidate construction actions. A trainable controller G_ϕ maps $\mathbf{w}$ to a feature gate and an optional action temperature for each action head or selection stage k:

$$\{(\mathbf{g}_k, \tau_k)\}_{k \in \mathcal{K}} = G_\phi(\mathbf{w}), \qquad \tau_k > 0. \quad (7)$$

In the reaction-based building-block selector, $k = c, b$ denotes cluster selection and within-cluster building-block selection. In fragment generation, k indexes the graph-construction action heads.

3.1 PREFERENCE-DEPENDENT FEATURE SCORING

For head k, let $\mathbf{h}_k(s)$ be its state representation and f_k its action scorer. The controller produces a multiplicative gate:

$$\mathbf{g}_k(\mathbf{w}) = 2\sigma(A_k E(\mathbf{w}) + \mathbf{b}_k), \qquad q_k(s, \mathbf{w}) = f_k(\mathbf{g}_k(\mathbf{w}) \odot \mathbf{h}_k(s)). \quad (8)$$

The fragment implementation initializes the gate at one using zero output weights and biases, so conditioning starts from the shared backbone policy. Feature dimensions are learned; they are not assumed to correspond one-to-one to molecular objectives.

For a normalized similarity scorer, let $\mathbf{e}_k(a)$ be the embedding of candidate action a. Following the reaction building-block formulation, the gate can rescale both representations after L_2 normalization:

$$\widetilde{\mathbf{h}}_k = \mathbf{g}_k(\mathbf{w}) \odot \frac{\mathbf{h}_k(s)}{\|\mathbf{h}_k(s)\|_2}, \qquad \widetilde{\mathbf{e}}_k = \mathbf{g}_k(\mathbf{w}) \odot \frac{\mathbf{e}_k(a)}{\|\mathbf{e}_k(a)\|_2}. \quad (9)$$

Numerical normalization uses a positive denominator floor. The resulting dot-product compatibility score is

$$q_k(s, a, \mathbf{w}) = \widetilde{\mathbf{h}}_k^\top \widetilde{\mathbf{e}}_k = \sum_j g_{k,j}(\mathbf{w})^2 \widehat{h}_{k,j}(s) \widehat{e}_{k,j}(a), \quad (10)$$

where hats denote normalized features. The finite-library control uses this bilinear form; the fragment model uses gated inputs to its action MLPs, rather than assuming every action head is a similarity scorer.

3.2 ACTION TEMPERATURES AND PREFERENCE CONTROL

Given compatibility logits, a stage policy normalizes over valid actions:

$$P_k(a \mid s, \mathbf{w}) = \mathrm{softmax}_{a \in \mathcal{A}_k(s)} \left(q_k(s, a, \mathbf{w}) / \tau_k(\mathbf{w}) \right). \quad (11)$$

Invalid actions are masked. In the fragment implementation, logits from all action types are normalized by the graph-action categorical distribution; type-specific temperatures can therefore also change competition between action types. Its temperature heads use softplus outputs clipped to $[0.1, 10]$ and initialized at one.

The feature gate changes which representation dimensions drive a decision, while temperature controls the concentration of the corresponding action distribution. Within a single head at a fixed state, a positive temperature preserves action rankings: $\mathrm{sign}[(q_a - q_b)/\tau] = \mathrm{sign}(q_a - q_b)$. Gating can change this sign by reweighting dimensions that favor different actions. This is a distinction

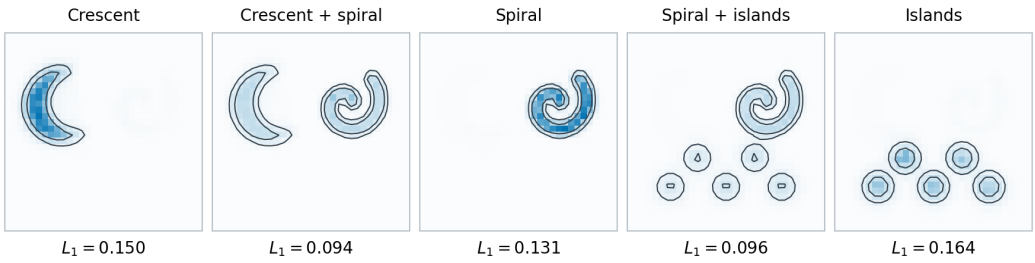

Figure 1: Toy example: preference control with one frozen PREFGATEGFN checkpoint (seed 0). Color shows learned terminal probability and contours show the exact target. Preferences move from crescent to spiral to islands through pairwise mixtures. The policy is not retrained between panels; L_1 errors enumerate all 1,024 states.

between mechanisms, not a guarantee of superior learned distributions; temperature alone can change terminal means. Learned action temperatures are internal policy parameters, distinct from the reward exponent β in Equation 1.

Gating thus changes the policy parameterization without itself changing the chosen reward or the valid construction space. The fragment model conditions every forward action head, including termination, and its state-flow estimator. Conditioning-architecture controls likewise match the action heads receiving preferences; the two building-block stages should not be read as a universal restriction on all implementations.

3.3 REWARD SCALING FOR COVERAGE

Dynamic normalization is a separate training component, not part of the gate definition. In the full-budget fragment setting it transforms objective scores using recent sampled statistics before scalarization (Appendix E.1). This changes the effective reward trade-offs and makes the training target nonstationary; it is not simply a change of evaluation units. We evaluate all checkpoints on frozen base scores.

3.4 CONTROLLED ALTERNATIVES

The molecular conditioning controls compare gates with a residual concatenation conditioner, $\mathbf{h} \leftarrow \mathbf{h} + F([\mathbf{h}; \mathbf{w}])$, and an HN-GFN-style preference-generated residual action map. Residual outputs start at zero. These are adaptations of the conditioning principles in Section 2.2, not claims of identical published implementations.

Within each conditioning-architecture control, arms share the backbone, state-flow or normalizer network, scalarization, loss, optimizer, update count, and preference tensors. Every stochastic action head receives the declared conditioning mechanism. The unconditioned arm omits preferences from policy logits but retains the conditional flow estimator. A 3×2 design crosses feature pathway (none, concatenation, gates) with temperature (off, on); the generated-head arm has temperature off. These controls use Dirichlet(1) preferences, without curriculum, renormalization, replay, EMA, or controller freezing. The separate full-budget training ablation retains EMA and the 480-update controller freeze while varying normalization and curriculum. Parameter counts may differ; our motivation does not depend on a capacity bottleneck.

4 EXPERIMENTS

4.1 TOY EXAMPLE: EXACT PREFERENCE CONTROL

We evaluate three complementary properties: candidate quality and Pareto coverage, response to requested trade-offs, and the separate effects of gates and temperature. Full-budget molecular experiments assess complete systems; inexpensive matched controls isolate components on a smaller scale.

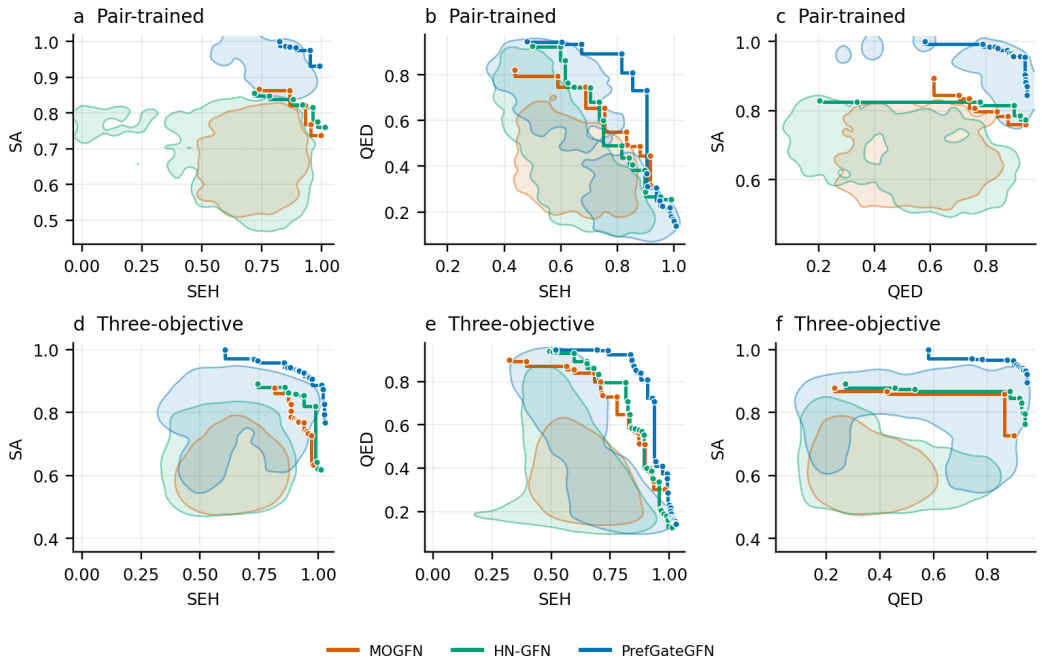

Figure 2: Main fragment-generation result. Top: policies trained on the indicated objective pair. Bottom: projections of policies trained jointly on SEH, QED, and SA. Lines and points show empirical two-dimensional non-dominated fronts (seed 0 for pair-trained policies; pooled seeds 0–2 for three-objective policies); shaded contours summarize smoothed 90% density regions. Fronts use all valid recorded draws, not the smoothed densities. Pair-trained and three-objective policies are separate experiments with 1,280 and 16,384 draws per seed, respectively; the sample budget and available seeds are matched across curves within each row. Compare methods within panels. The normalization–curriculum ablations are reported separately in Appendix E.1. Table 1 keeps the seed as the replicate.

This distinction allows us to report practical improvements without assigning every improvement to gating alone.

We begin with an enumerable 32×32 grid whose three equal-mass reward components form a crescent, a spiral, and disconnected islands. With linear scalarization and reward exponent one, the exact target is their preference-weighted mixture.

Figure 1 shows the learned distribution following changes in the requested shape, including transitions between connected and disconnected support. Across 66 simplex preferences and three seeds, terminal L_1 error is 0.0737 ± 0.0018 for PrefGateGFN, 0.0790 ± 0.0033 for concatenation with temperature, 0.0789 ± 0.0029 for MOGFN, and 0.0824 ± 0.0052 for HN-GFN. These modest differences provide a conditioning check, not a claim of molecular superiority. Appendix A gives training and checkpoint selection details, circle-mixture distributions, and the objective-count comparison.

4.2 Fragment generation: candidate quality and Pareto coverage

The fragment task scores SEH proxy activity, QED, and synthetic-accessibility desirability (SA), all oriented so that larger values are better (Bengio et al., 2021; Bickerton et al., 2012; Ertl & Schuffenhauer, 2009). SEH is the predictor output divided by eight; SA is $(10 - \mathrm{SA}_{\mathrm{raw}})/9$. We compare MOGFN, HN-GFN, and PrefGateGFN on each objective pair and on the joint three-objective task. The full-budget runs use 15,000 training updates with three seeds for the three-objective comparison. That evaluation uses the same 128 preferences and 128 draws per preference for each method and seed.

Table 1: Full-budget SEH–QED–SA results: mean ± sample SD over three seeds. Higher is better. Task, update budget, and evaluation are shared; training settings are detailed in Appendix E.

Method	Top-10 reward	Top-10 diversity	Hypervolume
MOGFN	0.659 ± 0.003	0.595 ± 0.007	0.615 ± 0.004
HN-GFN	0.729 ± 0.005	0.596 ± 0.009	0.669 ± 0.017
PREFGATEGFN	$\mathbf{0.829} \pm 0.022$	$\mathbf{0.648} \pm 0.025$	$\mathbf{0.809} \pm 0.070$

PREFGATEGFN extends the empirical frontier toward high SEH and high SA, while retaining high-QED candidates (Figure 2). This pattern appears in pair-trained policies and in projections of the three-objective policy. The density contours supply context: a frontier describes exceptional samples, not necessarily a typical draw. These are pairwise projections, not a claim that each displayed point is jointly optimal in all three dimensions.

Among the three systems in Table 1, PREFGATEGFN has the highest mean on all three metrics (Table 1). Its mean hypervolume is 0.140 above HN-GFN and 0.194 above MOGFN, accompanied by higher top-10 diversity rather than a reward-only improvement. These are complete-system results: the PREFGATEGFN runs include scheduled preferences, dynamic objective normalization, and an initial controller freeze; comparison runs use static objectives and Dirichlet preferences. The table therefore supports the performance of the evaluated system, not an architecture-only effect size. We isolate components in Section C.

Coverage and control are complementary. The full-budget normalization–curriculum ablation keeps the gated architecture fixed (Appendix E.1). Normalization alone gives hypervolume 0.885 ± 0.026, compared with 0.668 ± 0.012 with both components off and 0.681 ± 0.023 with curriculum alone. Thus the broader empirical frontier depends substantially on reward normalization, not on gating alone; adding the tested curriculum does not improve these means. All metrics use frozen base scores, so the gain is not an artifact of plotting normalized training rewards. Gating addresses a different requirement: conditioning action selection on user priorities. The next analysis tests that control directly for the original both-on system; it does not establish preference responsiveness of the new normalization-only checkpoints.

4.3 DOES CHANGING THE PREFERENCE CHANGE THE MOLECULES?

For a frozen checkpoint, let $\mu_i(\mathbf{w}) = \mathbb{E}_{x \sim p_\theta(\cdot|\mathbf{w})}[r_i(x)]$. We change the preference while holding the trained policy fixed and measure

$$\Delta_i = \mu_i(\mathbf{w}_{\text{high}}) - \mu_i(\mathbf{w}_{\text{low}}). \tag{12}$$

This directly tests the conditioning interface; aggregate hypervolume does not. Intermediate preferences and full distributions reveal whether an endpoint difference conceals a flat or non-monotone response.

On the full-budget SA-to-SEH edge (SEH weight zero to one), SEH gains are 0.007 ± 0.007 for MOGFN, 0.332 ± 0.021 for HN-GFN, and 0.179 ± 0.007 for PREFGATEGFN. On SA-to-QED, QED gains are 0.034 ± 0.021, 0.512 ± 0.003, and 0.461 ± 0.126, respectively. Thus PREFGATEGFN changes molecular outcomes in the requested direction on both edges and has substantially greater measured response than MOGFN in this setting. HN-GFN has the larger mean endpoint response: our advantage is not a uniformly steeper response curve, but the combination of useful preference response and stronger aggregate quality and coverage. Appendix E shows the distributions and full sweeps.

4.4 SYNTHESIS-AWARE GENERATION WITH MANY OBJECTIVES

We evaluate ADRB2, ALDH1, ESR agonist, ESR antagonist, and FEN1. Objective panels contain 5, 9, or 14 desirabilities: docking, one joint physicochemical criterion, and respectively 3, 7, or 12 ADMET endpoints (Swanson et al., 2024). Each target–system evaluation contains 1,000 draws from one training seed. We compare PREFGATEGFN, MOGFN-PC, and a concatenation control that retains the Huber/EMA configuration used in the gating comparison. HN-GFN is also evaluated on the 14-objective task.

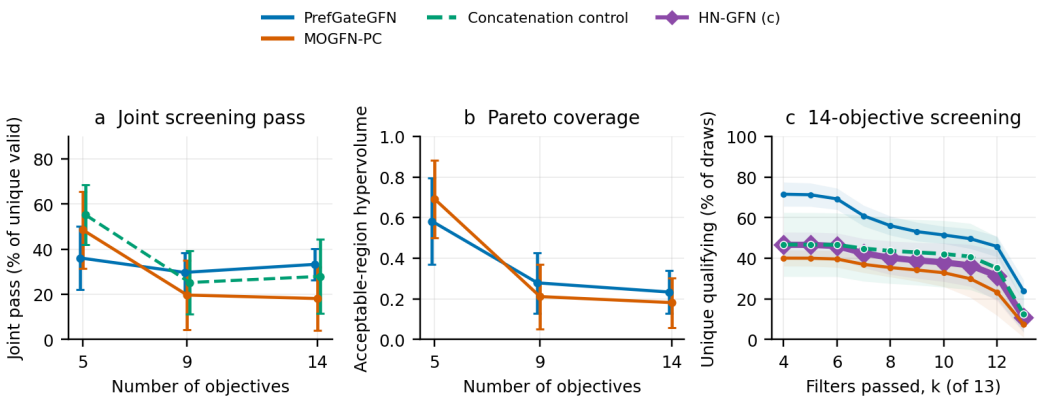

Figure 3: Synthesis-aware complete-system results across five targets. (a) Joint physicochemical and ADMET pass rate among unique valid molecules. (b) Acceptable-region hypervolume, including docking. (c) Unique molecules meeting at least k of the 13 non-docking filters in the 14-objective task, divided by the original 1,000-draw budget. HN-GFN appears in panel (c) only because its five-target evaluation covers the 14-objective task. Bars and bands are sample SD across targets, not training seeds. Training differences are in Appendix D.

The ordering changes as more objectives are included (Figure 3). PREFGATEGFN is not the best system on the five-objective panel, but attains the highest mean joint pass rate and acceptable-region hypervolume on the 14-objective panel: 33.3% and 0.234, versus 18.1% and 0.182 for MOGFN-PC. The concatenation control reaches a 28.0% joint pass rate. These outcomes support the practical value of PREFGATEGFN for this larger conjunction of requirements, without implying that every added objective necessarily favors gating.

Panel (c) also penalizes repeated molecules: PREFGATEGFN supplies 23.9 unique all-13-filter molecules per 100 draws, compared with 12.5 for concatenation control, 10.8 for HN-GFN, and 7.5 for MOGFN-PC. Adding docking score ≤ -8 kcal/mol gives 133, 102, 94, and 69 unique qualifying molecules per 1,000 draws, respectively (target means rounded to whole molecules). The calculations answer different questions: non-docking screening coverage and docking-filtered candidate yield.

The 14-objective yield gap has two measured components. Per 1,000 draws, PREFGATEGFN produces a mean of 715 unique valid molecules across targets, versus 467 for concatenation, 466 for HN-GFN, and 401 for MOGFN-PC. Among those unique molecules, the mean fractions passing all thirteen filters are 33.3%, 28.0%, 22.4%, and 18.1%, respectively. Thus panel (c) reflects both more distinct proposals and a larger fraction surviving the conjunction; these descriptive counts do not identify which training component caused the difference.

The full reaction comparison differs in controller updating and preference schedules across systems.

4.5 SIMULTANEOUS DILI AND BIOAVAILABILITY REQUIREMENTS

We additionally inspect a biologically relevant pair of joint design considerations: low predicted drug-induced liver injury (DILI) liability and high predicted oral bioavailability. Define the direction-aligned scores $u_{\text{DILI}} = 1 - p_{\text{DILI}}$ and $u_{\text{BA}} = p_{\text{bioavailability}}$.

PREFGATEGFN reaches the stringent joint region on every target (Figure 4). Its distinct joint-pass counts are 12, 45, 12, 46, and 34 for ADRB2, ALDH1, ESR agonist, ESR antagonist, and FEN1, respectively. The corresponding MOGFN-PC counts are 2, 0, 0, 0, and 5. HN-GFN yields 0, 0, 5, 11, and 2; the concatenation control yields 4, 0, 1, 6, and 0. Averaged across targets, the counts are 29.8 for PREFGATEGFN, 3.6 for HN-GFN, 2.2 for concatenation, and 1.4 for MOGFN-PC.

The stringent pairwise threshold sharpens this distinction: averaged across targets, 4.1% of PREFGATEGFN's unique valid molecules pass both requirements, versus 0.7% for HN-GFN, 0.5% for concatenation, and 0.4% for MOGFN-PC. One plausible explanation is the reaction gate in Equation 10: preferences directly reweight the shared state–action feature matches used to rank building

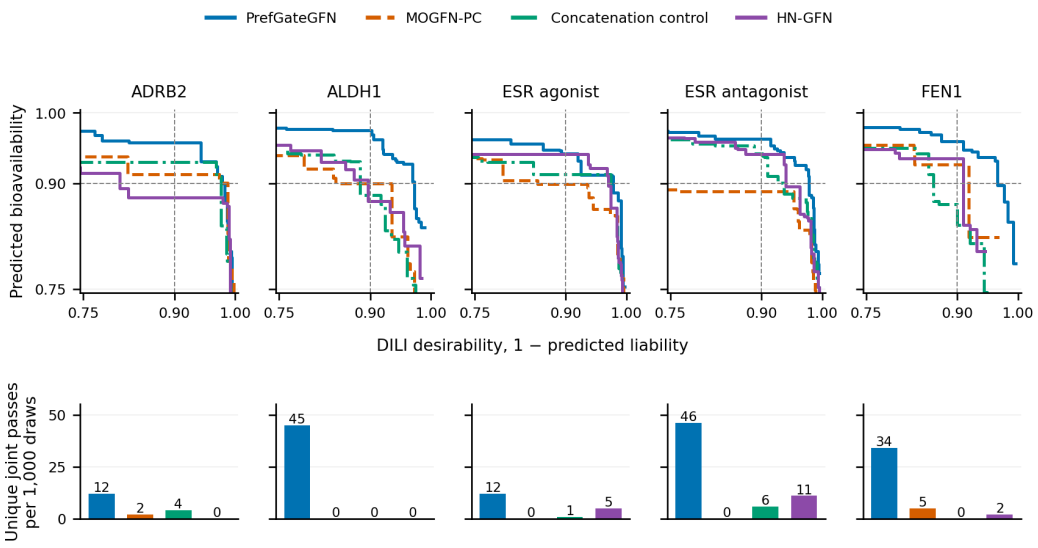

Figure 4: **Many-objective generation reaches the joint DILI–bioavailability region.** Top: empirical pairwise Pareto fronts of the fourteen-objective systems on five targets; dashed reference lines indicate 0.9 on both axes. Bottom: distinct valid molecules satisfying both $u_{\mathrm{DILI}} \geq 0.9$ and $u_{\mathrm{BA}} \geq 0.9$, per 1,000 original draws. Invalids and duplicates consume the budget. Counts use all recorded molecules, not a smoothed density or selected frontier points. HN-GFN and the concatenation control use the same five targets and draw budget. This pairwise screen does not impose the remaining twelve objectives. Each system has one training seed per target.

blocks. If useful trade-offs mostly reweight a common feature basis, this gives a compact way to favor jointly desirable construction choices while retaining shared action representations. Concatenation must learn those interactions through its residual feature transform; HN-GFN instead generates a more flexible dense action map, which may require more finite-budget evidence to estimate. This is an inductive-bias hypothesis, not a causal attribution or a claim that dense maps are intrinsically worse. The reaction systems differ in controller updates and preference schedules, and each target has one training seed; matched architecture controls are needed to isolate gating.

Requiring the physicochemical criterion and all twelve standard ADMET filters in addition to these two stringent thresholds leaves 20.4 distinct PREFGATEGFN candidates per 1,000 draws on average. Adding docking score ≤ -8 kcal/mol leaves 12.6, versus 0.2 for MOGFN-PC. This most stringent intersection is target-dependent: PREFGATEGFN yields 0, 33, 3, 27, and 0 candidates across the five targets. It is therefore not a universal solution to every stringent constraint, even though the standard fourteen-objective screen yields qualifying candidates on every target.

USE OF AI ASSISTANCE

Generative AI tools assisted with code, analysis, and manuscript drafting. Experimental values are computed from recorded runs or stored molecular evaluations. Scientific interpretation and final author approval remain the authors' responsibility.

REFERENCES

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27381–27394. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e614f646836aaed9f89ce58e837e2310-Paper.pdf.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023. URL https://jmlr.org/papers/v24/22-0364.html.

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

Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio, and Pietro Liò. SynFlowNet: Design of Diverse and Novel Molecules with Synthesis Constraints. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7495fa446f10e9edef6e47b2d327596e-Abstract-Conference.html.

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

Moksh Jain, Sharath Chandra Raparthy, Alex Hernández-García, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective GFlowNets. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14631–14653. PMLR, 2023. URL https://proceedings.mlr.press/v202/jain23a.html.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4849–4859. PMLR, 2020. URL https://proceedings.mlr.press/v119/jin20b.html.

Kanika Madan, Jarrid Rector-Brooks, Maksym Korablyov, Emmanuel Bengio, Moksh Jain, Andrei Cristian Nica, Tom Bosc, Yoshua Bengio, and Nikolay Malkin. Learning GFlowNets from partial episodes for improved convergence and stability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23467–23483. PMLR, 2023. URL https://proceedings.mlr.press/v202/madan23a.html.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlowNets. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0431.

Christos A. Nicolaou and Nathan Brown. Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies*, 10(3):e427–e435, 2013. doi: 10.1016/j.ddtec.2013.02.001.

Matthew D. Segall. Multi-parameter optimization: Identifying high quality compounds with a balance of properties. *Current Pharmaceutical Design*, 18(9):1292–1310, 2012. doi: 10.2174/138161212799436430.

Matthew D. Segall. Advances in multiparameter optimization methods for de novo drug design. *Expert Opinion on Drug Discovery*, 9(7):803–817, 2014. doi: 10.1517/17460441.2014.913565.

Seonghwan Seo, Minsu Kim, Tony Shen, Martin Ester, Jinkyoo Park, Sungsoo Ahn, and Woo Youn Kim. Generative flows on synthetic pathway for drug design. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pB1XSj2y4X.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C. Wu, Rabindra V. Shivnaraine, and James Zou. ADMET-AI: A machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024. doi: 10.1093/bioinformatics/btae416.

Yuejiang Yu, Chun Cai, Jiayue Wang, Zonghua Bo, Zhengdan Zhu, and Hang Zheng. Uni-Dock: GPU-accelerated docking enables ultralarge virtual screening. *Journal of Chemical Theory and Computation*, 19(11):3336–3345, 2023. doi: 10.1021/acs.jctc.2c01145.

Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, Chang-Yu Hsieh, Tingjun Hou, and Jian Wu. Sample-efficient multi-objective molecular optimization with GFlowNets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A EXACT-GRID PROTOCOL

The shape-atlas models were trained for 20,000 updates, with Adam learning rate 10^{-3}, batch size 128, width-64 two-layer state networks, SubTB geometric weighting parameter 2, random-action probability 0.05, and replay capacity 10,000. The preference encoder width is 16 for the gated and concatenation implementations; the HN-GFN preference network uses width 32. Seeds are 0–2. The saved checkpoint is selected by the training implementation's lowest validation L_1 at 100-update intervals, not necessarily the final update. For example, the displayed gated seed-0 checkpoint is from update 19,100. Evaluation reuses the stored checkpoints and enumerates terminal probabilities at a step-0.1 simplex grid (66 preferences). The displayed shape path includes vertices and should not be read as the average error on that grid.

For a directed path from component a to b, we normalize the expected b-reward by its exact values at the two endpoints. Equal component mass makes the exact response linear. Response RMSE is evaluated along all directed component pairs. This calibration is specific to the synthetic reward construction and is not imposed on molecular data. The modest shape-atlas differences measure distribution fitting. The grid implementations also differ in their flow-head parameterizations; the molecular controls provide the cleaner component comparison.

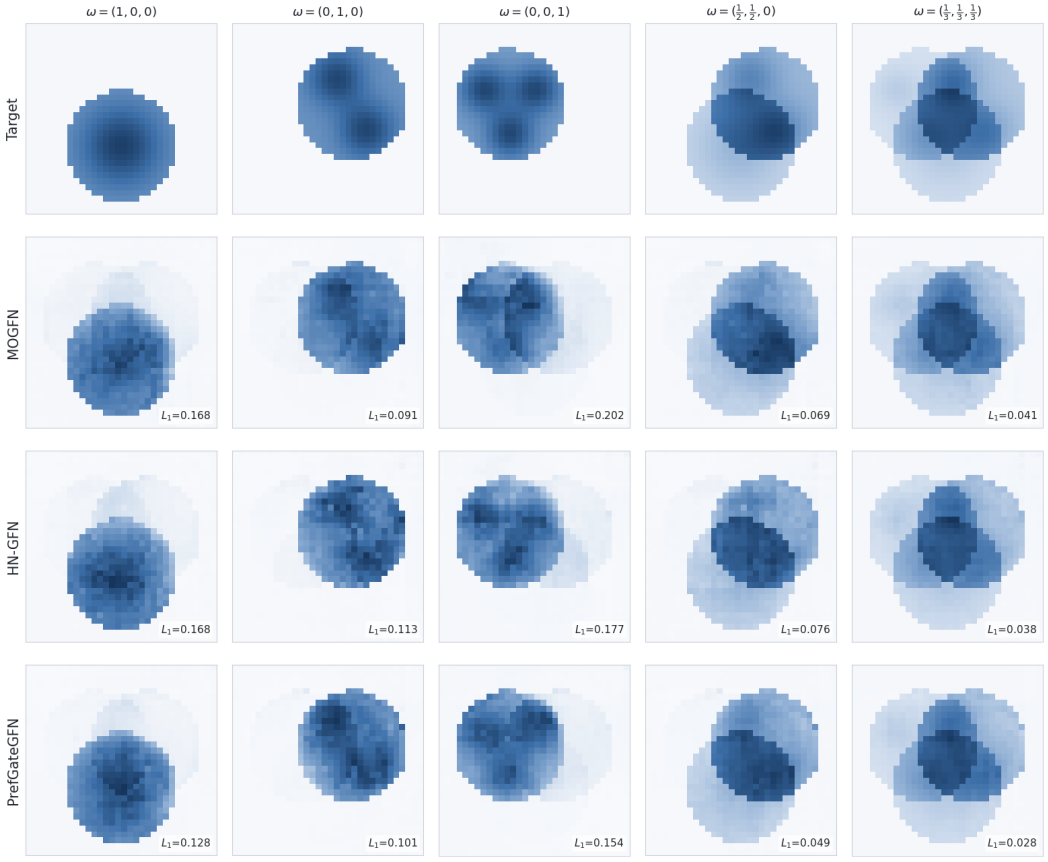

Figure 5: Circle-mixture preference intervention. Columns change the preference; rows compare the exact target with MOGFN, HN-GFN, and PREFGATEGFN at seed 0. Annotated L_1 values compare complete terminal distributions, not sampled histograms. This illustration uses the circle reward construction; the scale-normalized objective ladder in Table 2 is a separate experiment and should not be numerically conflated with these cells.

PREFGATEGFN is close to HN-GFN on the four- through eight-objective panels and has lower mean error than MOGFN there; HN-GFN is stronger on the two- and three-objective panels. Adding objectives also changes reward geometry, so this is a family of tasks, not an isolated dimensionality

Table 2: Objective-count ladder: exact terminal L_1, mean $\pm$ sample SD over three seeds; lower is better. Bold marks the lowest mean in each row, not statistical significance. Each component of the `nmix` panels is min–max normalized on the grid. No one-objective run is present in the recorded comparison (—); it is not assigned an inferred score.

Objectives	MOGFN	HN-GFN	PREFGATEGFN
1	—	—	—
2	0.0229 ± 0.0010	$\mathbf{0.0191} \pm 0.0032$	0.0265 ± 0.0091
3	0.0300 ± 0.0001	$\mathbf{0.0249} \pm 0.0015$	0.0324 ± 0.0069
4	0.0432 ± 0.0027	$\mathbf{0.0335} \pm 0.0008$	0.0336 ± 0.0071
5	0.0488 ± 0.0020	0.0357 ± 0.0020	$\mathbf{0.0349} \pm 0.0004$
6	0.0869 ± 0.0056	$\mathbf{0.0647} \pm 0.0021$	0.0649 ± 0.0003
7	0.1441 ± 0.0078	$\mathbf{0.1060} \pm 0.0035$	0.1085 ± 0.0010
8	0.2184 ± 0.0108	0.1546 ± 0.0029	$\mathbf{0.1525} \pm 0.0052$

intervention. With one objective there is no preference trade-off; the unmeasured row is retained to make the requested range explicit.

B CONTROLLED FRAGMENT PROTOCOL

The environment and SEH predictor are inherited from the fragment GFlowNet implementation (Bengio et al., 2021). SEH is the proxy output divided by 8; QED is the drug-likeness score (Bickerton et al., 2012); SA desirability is $(10 - \mathrm{SA}_{\mathrm{raw}})/9$ (Ertl & Schuffenhauer, 2009). The fragment policy uses the upstream GFlowNet implementation with PyTorch and PyTorch Geometric; RDKit supplies molecular parsing and QED scoring. SEH is not assumed to be bounded by one. The scalar reward is the weighted sum, raised to the fixed reward exponent 16.

All seven arms use a two-layer graph transformer of width 64 and two attention heads. The backward policy is uniform. The loss is SubTB with squared residuals; Adam uses learning rate 5×10^{-4} and weight decay 10^{-8}, with the same upstream learning-rate decay and gradient clipping in every arm. Each of 400 updates samples 32 trajectories, with random-action probability 0.05 during training and zero during evaluation. There is no EMA, replay, curriculum, dynamic reward normalization, or freezing. The conditional state-flow estimator is an identical width-64 network in every arm, including the policy-unconditioned control.

Residual concatenation modulates features immediately before every action head. The HN-GFN adaptation generates a residual linear map and bias for each action type. Gates multiply the corresponding action features. All use width-64 preference networks. The temperature factor has one output per action type. Common backbone and flow tensors are checked for identical initialization across arms within a seed; declared neutral residuals and gates recover the same policy. Parameter counts and resolved configurations are stored with each run.

The six evaluation vectors, in SEH–QED–SA order, are $(.2, 0, .8)$, $(.5, 0, .5)$, $(.8, 0, .2)$, $(0, .2, .8)$, $(0, .5, .5)$, and $(0, .8, .2)$. Sampling seeds depend on training seed and checkpoint, not architecture. Common random seeds reduce a source of variation but do not imply identical sample trajectories between models. The final contrast is computed for each training seed before aggregation. With only 64 draws per preference, small changes can be dominated by sampling noise.

All 24,192 recorded evaluation draws in this pilot were valid. Training and evaluation across the 21 runs took approximately 1.50 GPU-hours, excluding startup and the separately recorded preflight and interrupted timing attempts.

C SEPARATING FEATURE GATES FROM TEMPERATURE

A fixed library of 512 reaction-generated ADRB2 molecules isolates conditioning components without new docking calls. Candidates are deduplicated and selected by a score-independent hash from three generators; a one-step GFlowNet selects among them and their recorded routes. Rewards are docking desirability, one minus predicted hERG liability, and predicted bioavailability. All seven

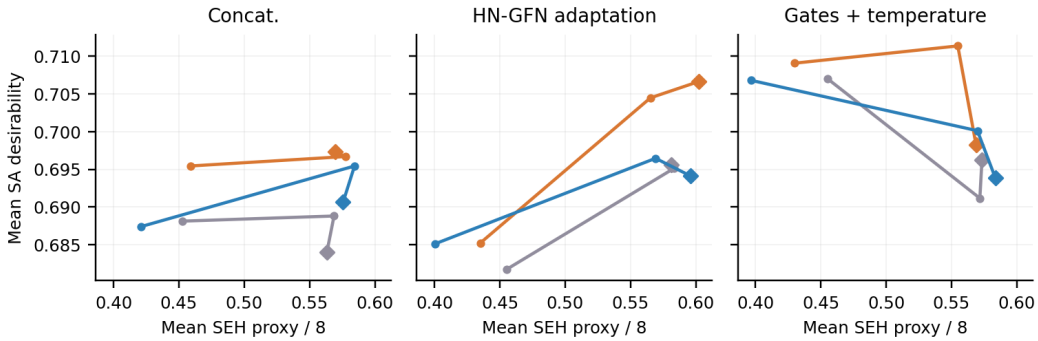

Figure 6: Preference-specific centroid paths over updates 50, 200, and 400. Each point averages the three training seeds; the diamond marks update 400. Orange, grey, and blue denote SEH weights 0.2, 0.5, and 0.8, respectively, with QED weight zero. These paths use the measured checkpoints, not interpolated or synthetic training data.

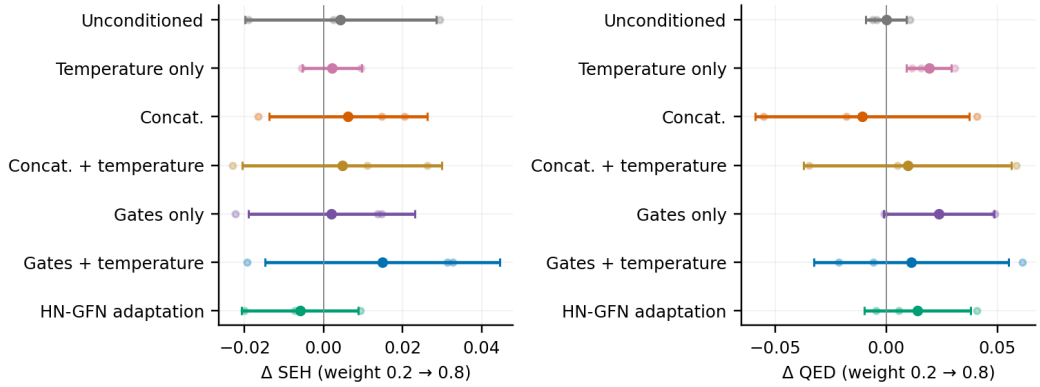

Figure 7: All seven fragment arms at update 400. Endpoint response is the change in the requested property as its weight increases from 0.2 to 0.8 along the edge with SA. Error bars show the SD across three training seeds; faint dots are the individual seed estimates. Both sides of zero are retained.

Table 3: Controlled molecular studies (mean ± SD). Fragment responses use three training seeds at update 400; the exact library error uses five seeds at update 800. Each study matches all non-architecture settings within its own arms. Bold marks the largest mean response or lowest mean error, not statistical significance. Larger response is not automatically better target calibration.

Conditioning	Fragment Δ SEH	Fragment Δ QED	Library L_1
Unconditioned	0.004 ± 0.024	-0.000 ± 0.009	0.628 ± 0.024
Temperature only	0.002 ± 0.008	0.019 ± 0.010	0.511 ± 0.011
Concatenation	0.006 ± 0.020	-0.011 ± 0.048	0.281 ± 0.015
Concatenation + temperature	0.005 ± 0.025	0.010 ± 0.047	0.270 ± 0.010
Gates only	0.002 ± 0.021	$\mathbf{0.024} \pm 0.025$	0.292 ± 0.015
Gates + temperature	$\mathbf{0.015} \pm 0.030$	0.011 ± 0.044	$\mathbf{0.262} \pm 0.014$
HN-GFN adaptation	-0.006 ± 0.015	0.014 ± 0.024	0.277 ± 0.034

arms share geometric scalarization, 800 TB updates, five seeds, preference banks, molecular encoder, and normalizer. There is no curriculum, EMA, replay, or freezing. We enumerate the exact target at 231 preferences and measure $L_1 = \sum_x |p_\theta(x \mid \mathbf{w}) - p^*(x \mid \mathbf{w})|$.

Gates alone reduce mean error from 0.628 to 0.292; temperature alone reduces it to 0.511; together they reach 0.262 ± 0.014. Adding temperature to gates improves mean error by 0.030. Concatenation with temperature reaches 0.270 ± 0.010, and the HN-GFN-style arm reaches 0.277 ± 0.034. These results support separate contributions from feature conditioning and temperature, with close means among expressive conditional models. The gate-by-temperature interaction in L_1 is $+0.087$: the improvements are not additive and do not establish synergy. The protocol below gives the complete experimental details, and Figure 8 shows the exact-distribution errors and response curves.

As a second control, all seven arms are trained in the de novo fragment environment for 400 updates at three seeds, with shared preference banks, static rewards, identical optimization, and no curriculum or freezing. Appendix B reports property-centroid trajectories at updates 50, 200, and 400 and all endpoint contrasts. Responses are small relative to seed variation: this short-budget experiment is inconclusive about architecture ranking at convergence. It neither replaces the full-budget results nor establishes that gating resolves learning imbalance. All 24,192 evaluation draws were valid; training and evaluation used approximately 1.50 GPU-hours.

C.1 Finite reaction-molecule library protocol

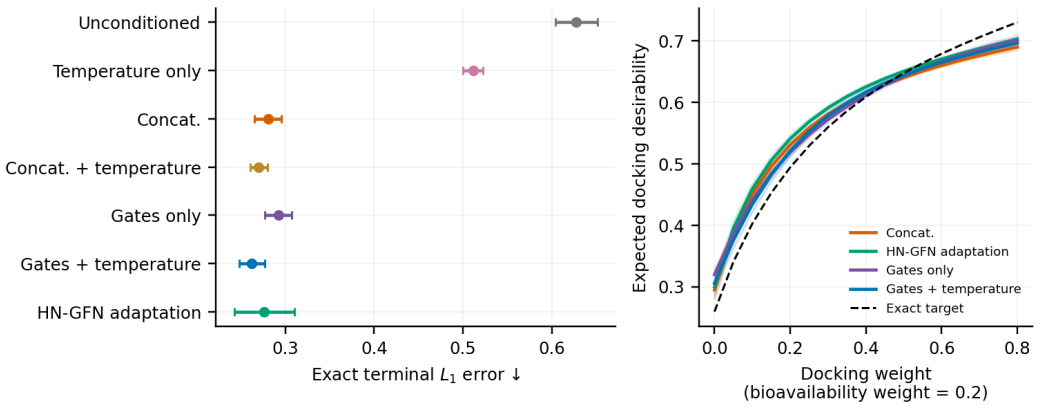

Figure 8: Matched molecular-library ablation. Left: terminal L_1, mean ± SD over five seeds. Right: docking response at bioavailability weight 0.2, with the remaining weight transferred from hERG to docking; the dashed curve is the exact target. This isolates components on a fixed molecular support, not unrestricted reaction generation.

The source pool comprises 1,000 ADRB2 draws from each of three existing systems: PREF-GATEGFN, concatenation with Huber/EMA, and MOGFN-PC. Valid molecules are deduplicated

by canonical SMILES in that fixed source order, retaining the first stored score and route. SHA-256 of canonical SMILES determines the 512 retained candidates; no reward threshold enters selection. Source files and the realized library are hashed. This pool inherits the coverage of those generators and is not an unbiased sample of chemical space.

Each candidate is encoded by a radius-2, 256-bit Morgan fingerprint and a two-layer width-32 MLP. A learned state vector scores the resulting embedding with a shared candidate bias. Gates multiply the bilinear coordinates as in Equation 10. The concatenation arm produces a residual state vector from state and preference. The HN-GFN adaptation generates a full 32×32 residual bilinear map. Temperature comes from a separate width-32 network and is initialized identically in all enabled cells. Unused preference modules in disabled cells do not supply a policy signal.

The docking desirability is $\sigma((\max(-d, 0) - 8)/1.5)$ for stored docking score d. The other desirabilities are $1 - p_{\text{hERG}}$ and $p_{\text{bioavailability}}$ from stored ADMET-AI predictions (Swanson et al., 2024). Values are clipped below at 10^{-4}. The forward trajectory has one action and the backward probability is one. Every training update averages the squared TB residual over all 512 trajectories and 32 sampled preferences. Adam uses learning rate 0.002 and weight decay 10^{-6} for 800 updates, seeds 0–4; no EMA, replay, curriculum, or freezing is used. Evaluation enumerates the step-0.05 simplex (231 preferences), including vertices. The same support is used for fitting and evaluation; no generalization to new molecules is claimed.

D EXISTING REACTION-GENERATION RECORDS

The five targets are ADRB2, ALDH1, ESR agonist, ESR antagonist, and FEN1. For each target and system, the recorded evaluation contains 1,000 draws from training seed 0. A strict hit is a unique valid canonical molecule passing the stored physicochemical criterion, every one of the 12 stored ADMET criteria, and docking score at most -8 kcal/mol. We count each canonical molecule once and divide by the original 1,000 draws, retaining invalids and duplicates in the budget denominator.

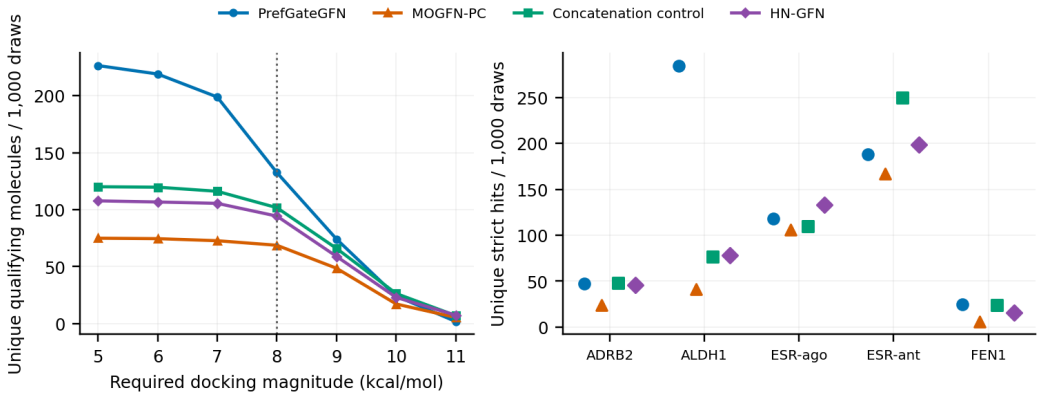

Figure 9: Descriptive reanalysis of existing full reaction-generation systems. Left: unique qualifying yield versus the required docking magnitude, averaged equally over five targets. Right: the individual target counts at 8 kcal/mol. Each system has one training seed per target. Training recipes differ, so these points are not an architecture-controlled ablation.

Mean unique strict hits per 1,000 draws are 133 for PREFGATEGFN, 102 for concatenation with Huber/EMA, and 69 for MOGFN-PC. These are system-level outcomes. The gated controller was trained for about 950 of 3,000 updates, while the MOGFN-PC conditioner was not placed in the corresponding frozen parameter group. The stored preference schedule also behaved differently across systems. Loss and EMA differ between the original MOGFN-PC and Huber/EMA recipes. These confounds preclude attribution to gating, temperature, or curriculum.

A docking-success flag only records a returned favorable-sign score; it is not equivalent to passing the -8 threshold. The explicit threshold and deduplicated budget denominator explain why these

numbers differ from duplicate-counting pass rates. Computational docking and ADMET predictions remain surrogate endpoints, not measurements of binding or clinical safety.

E FULL-BUDGET FRAGMENT PROTOCOL AND PREFERENCE DIAGNOSTICS

The main fragment results use 15,000-update three-objective checkpoints (three seeds per system), evaluated on 128 preferences with 128 draws each. The pair-trained panels use seed 0 for every method, with 1,280 draws each: additional PREFGATEGFN seeds are not pooled into those panels because matching baseline seeds are unavailable. This prevents unequal sampling opportunities from favoring its empirical frontier.

The three-objective PREFGATEGFN schedule begins with 480 updates of uniform-preference backbone warm-up, followed by 240 updates for each single-objective focus phase, 960 updates with concentrated Dirichlet(10) preferences, and then Dirichlet(1) preferences through update 15,000. Each update uses 64 trajectories (960,000 overall). During backbone warm-up, the preference MLP, gate heads, temperature heads, and flow gate are frozen; all are unfrozen after warm-up. The operational rationale is to establish shared molecular features before adapting them to individual objectives. Dynamic normalization uses a window of 10,000 values per objective, a 1,000-value burn-in, and updates after 100 additional values. It maps r_i to $\max(6(r_i - \mu_i)/\sigma_i + 5, 10^{-5})$ when $\sigma_i > 0$. These training transformations are not used as the property axes in the evaluation figures. This schedule is a training choice, not an intrinsic requirement of gating; the full-budget ablation below tests it separately from normalization. MOGFN and HN-GFN comparison runs use static objectives and Dirichlet preferences. Accordingly, we report their full-budget results as a comparison of complete systems and disable curriculum and freezing uniformly in the short-budget conditioning controls. The full-budget training ablation instead retains the same 480-update freeze in every arm to avoid changing a third variable.

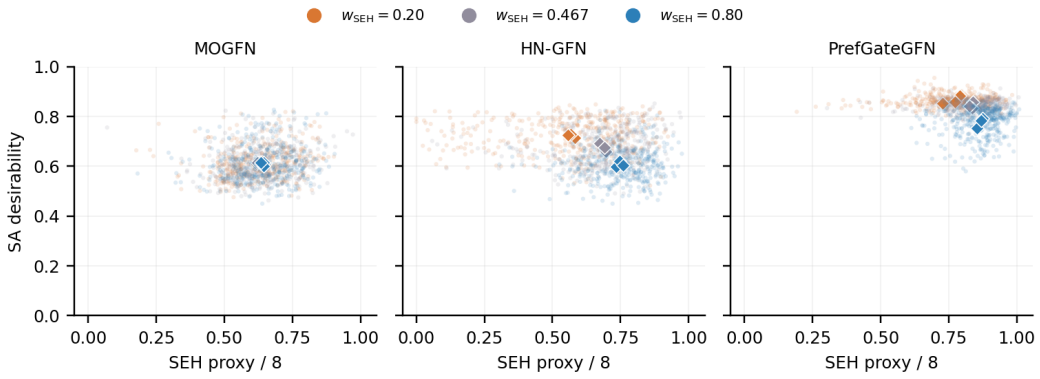

Figure 10: Full-budget final-checkpoint fragment distributions under direct preference intervention. Each color uses 128 draws per training seed; diamonds are individual seed centroids. The middle preference is the available near-balanced point $(7/15, 0, 8/15)$, not $(.5, 0, .5)$. Differences between systems cannot be assigned solely to architecture.

Top-10 reward is computed separately within each evaluation preference and then averaged across preferences, not maximized over preferences. Top-10 diversity is the evaluator's Tanimoto-distance summary of those high-reward molecules. Hypervolume uses all non-dominated objective vectors with a zero reference. SEH is not clipped at one, so this hypervolume is not intrinsically bounded by one. Figure 2 uses an exact two-dimensional front computation; smoothing affects only the background density contours.

E.1 SEPARATING DYNAMIC NORMALIZATION FROM PREFERENCE CURRICULUM

Drug discovery requires both a useful range of candidate profiles and a way to select among those profiles. This experiment tests the training components contributing to the first requirement; it does not replace the preference-intervention tests of the second. We complete a 2×2 comparison

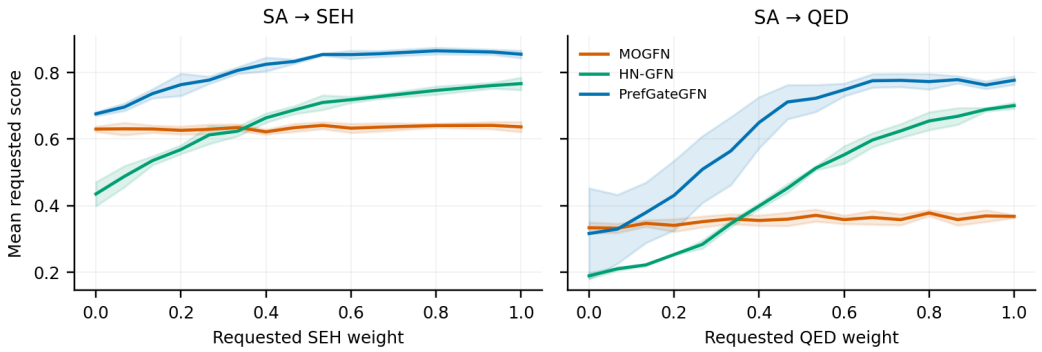

Figure 11: Full-budget fragment edge sweeps. Curves average all 128 draws per preference within each seed, then average seeds; bands are seed SD. The endpoint SEH gains are 0.007 ± 0.007, 0.332 ± 0.021, and 0.179 ± 0.007 for MOGFN, HN-GFN, and PREFGATEGFN. Responsiveness is not uniformly greatest for the gated system. These are our own checkpoint diagnostics, not comparisons to published reproduction scores.

of dynamic reward normalization and preference curriculum within the same PREFGATEGFN architecture. The both-on arm uses the original three-seed full-budget runs; the other three arms are new runs of the release implementation. All arms retain gates, learned temperatures, the six-layer width-128 backbone, SubTB loss, optimizer, reward exponent 32, sampling EMA 0.99, exploration probability 0.05, and the 480-update controller freeze. They use seeds 0–2, 15,000 updates of 64 trajectories, final checkpoints, and the same 128 evaluation preferences and 128 draws per preference. Release model-source identity and resolved training configurations are checked against the original runs.

With curriculum off, preferences follow Dirichlet(1) throughout training, including the controller-freeze period. With normalization off, the rewards remain SEH/8, QED, and $(10 - \mathrm{SA}_{\mathrm{raw}})/9$. When enabled, each component uses the schedule or running-statistics transformation specified above. Evaluation always uses frozen base scores; no dynamic normalization is applied to the recorded objective vectors or metrics. Thus the comparison measures changes in generated molecular profiles rather than a change of evaluation scale.

Table 4: Full-budget PREFGATEGFN training ablation. Mean $\pm$ sample SD over three training seeds; higher is better. Bold indicates the largest mean per metric, not a significance test. The both-on arm is the original configuration in Table 1.

Normalization	Curriculum	Top-10 reward	Top-10 diversity	Hypervolume
Off	Off	0.728 ± 0.001	0.594 ± 0.003	0.668 ± 0.012
Off	On	0.734 ± 0.008	0.603 ± 0.009	0.681 ± 0.023
On	Off	$\mathbf{0.844} \pm 0.014$	$\mathbf{0.679} \pm 0.018$	$\mathbf{0.885} \pm 0.026$
On	On	0.829 ± 0.022	0.648 ± 0.025	0.809 ± 0.070

Normalization has the larger observed effect on coverage in this experiment: enabling it without curriculum raises mean hypervolume from 0.668 to 0.885, whereas enabling curriculum without normalization raises it to 0.681. Normalization-only also has the highest mean reward and diversity. The both-on configuration remains competitive, but its mean hypervolume is lower than normalization-only; the results therefore do not support a claim that curriculum and normalization must be combined for the best outcome. The three-seed study supports normalization as a useful training component within the tested gated model, not as a uniquely gating-specific advantage: the same normalization ablation has not been run for MOGFN or HN-GFN. Nor does a broader frontier establish better preference control. Direct preference sweeps for these new ablation checkpoints remain to be evaluated; the full-budget response results above concern the original both-on system.

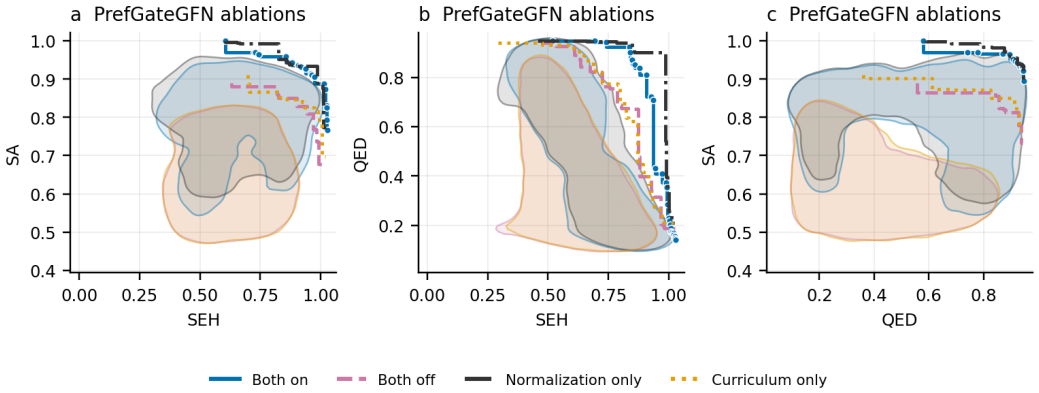

Figure 12: PREFGATEGFN normalization–curriculum ablations on SEH, QED, and SA. All configurations pool the same three seeds and 49,152 draws. Lines show exact empirical two-dimensional non-dominated fronts; shaded contours show smoothed 90% density regions. Blue: both on; dashed purple: both off; dash-dotted dark gray: normalization only; dotted gold: curriculum only. All axes use frozen base scores. These are projections, not a claim of dominance in every region of the three-dimensional objective space.

F REACTION SCREENING, HYPERVOLUME, AND COMPONENT COMPARISONS

F.1 SOFTWARE AND ENDPOINT DEFINITIONS

The reaction generator is implemented in RxnFlow using PyTorch for policy optimization and PyTorch Geometric for batched molecular graphs. RDKit parses and canonicalizes SMILES, applies reaction templates, and computes the physicochemical descriptors. The reward worker calls UniDock (Yu et al., 2023) through `unidock_tools` for docking and ADMET-AI for the twelve predicted endpoints. NumPy and pandas process evaluation records; Matplotlib renders the figures. The generator and reward worker use separate dependency environments because their chemistry and prediction packages have different requirements. These are predicted screening scores, not measured biological outcomes.

The five-objective panel contains docking, physicochemical suitability, hERG, PAMPA, and P-gp. The nine-objective panel adds BBB, Ames, DILI, and bioavailability. The fourteen-objective panel additionally includes CYP3A4, ClinTox, SR-MMP, NR-AR, and CYP2D6. ADMET outputs are from ADMET-AI (Swanson et al., 2024). Liability predictions are converted to desirabilities using $1 - p$; positively oriented predictions retain p. The physicochemical objective combines molecular weight, hydrogen-bond donors, topological polar surface area, and LogP. Joint screening uses the stored endpoint pass flags and physicochemical criterion. Each ADMET filter passes at desirability ≥ 0.5; missing or non-finite predictions fail the evaluation filter. These are computational screening definitions, not clinical acceptability thresholds.

Table 5: Predicted ADMET endpoints in the reaction objective panels. "First panel" gives the smallest panel containing that endpoint; larger panels retain all earlier endpoints. We maximize the positive predictions and minimize predicted liabilities.

Endpoint	ADMET-AI output	Direction	First panel
hERG inhibition	`hERG`	Lower	5
PAMPA permeability	`PAMPA_NCATS`	Higher	5
P-glycoprotein inhibition	`Pgp_Broccatelli`	Lower	5
Blood–brain barrier penetration	`BBB_Martins`	Higher	9
Ames mutagenicity	`AMES`	Lower	9
Drug-induced liver injury	`DILI`	Lower	9
Oral bioavailability	`Bioavailability_Ma`	Higher	9
CYP3A4 inhibition	`CYP3A4_Veith`	Lower	14
Clinical toxicity	`ClinTox`	Lower	14
Mitochondrial toxicity	`SR-MMP`	Lower	14
Androgen-receptor activity	`NR-AR`	Lower	14
CYP2D6 inhibition	`CYP2D6_Veith`	Lower	14

The training settings matter for interpretation. The PREFGATEGFN reaction controller is in the hypernetwork parameter group: its warm-up freeze ends at update 50 and its late backbone-only phase begins at update 1,000, giving about 950 controller updates out of 3,000. MOGFN-PC's conditioner was outside that frozen group and remained trainable throughout. Moreover, the intended preference curriculum was inactive in the recorded PREFGATEGFN runs, which used Dirichlet preferences, whereas other arms reached their scheduled fixed profiles. We describe this effective behavior rather than claiming identical schedules. Huber/EMA also distinguishes the concatenation control from the MOGFN-PC comparator. The plots measure the evaluated systems; these differences prevent their gaps from being interpreted as pure gating effects. All new molecular controls instead keep controllers trainable and disable curriculum uniformly.

F.2 JOINT SATISFACTION BEYOND MARGINAL PASS RATES

For target t, let q_{tj} be the fraction of draws passing ADMET endpoint j, and J_t the fraction passing all twelve. Independence predicts $I_t = \prod_{j=1}^{12} q_{tj}$. We compare this descriptive reference with J_t; the ratio J_t/I_t quantifies enrichment of the conjunction relative to the product of its marginal rates. A large ratio need not imply a high absolute yield, because its denominator can be small. Molecular

endpoints are biologically and statistically dependent, so this is not a null model of the chemistry or evidence that the architecture created those dependencies.

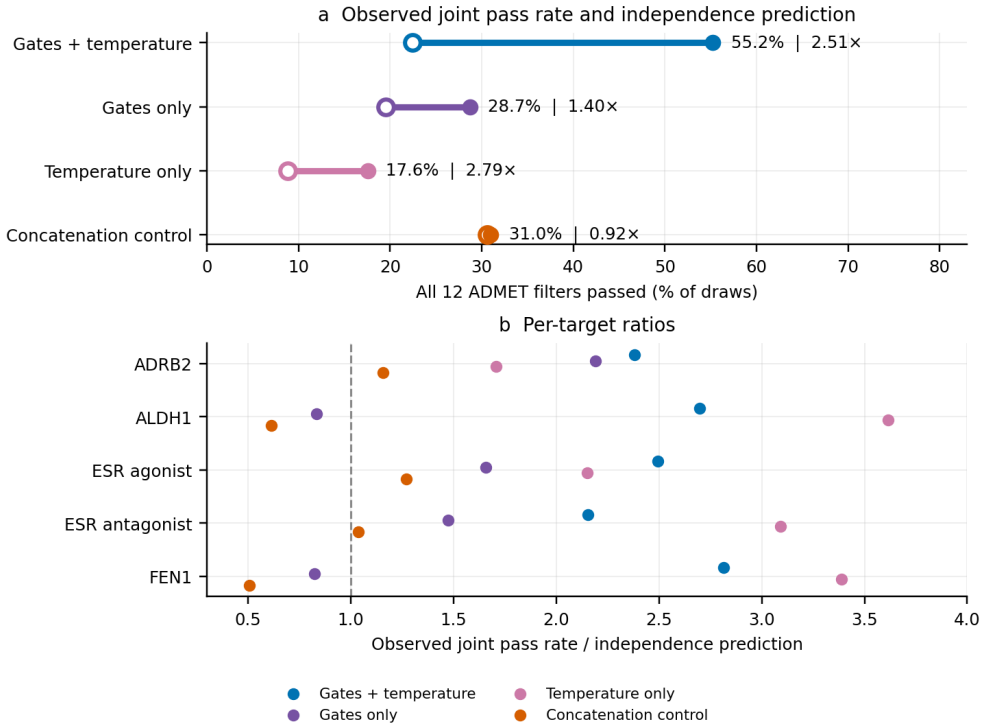

Figure 13: Reaction component comparison, 1,000 draws per target and arm. (a) Open circles mark the mean independence prediction; filled circles mark the mean observed all-12-ADMET pass rate. Annotations give the observed rate and the mean of the five target-wise ratios J_t/I_t, not the ratio of the displayed means. (b) Individual target ratios. Both panels count draws, including duplicates; neither adds the physicochemical or docking filter. The concatenation control retains a preference pathway. Training differences make this a system-level component comparison, not the matched factorial in Figure 8.

The full PREFGATEGFN system has a 55.2% mean joint ADMET pass rate, compared with 28.7% for gates only, 17.6% for temperature only, and 31.0% for concatenation. Its mean target-wise enrichment is 2.51. Temperature only has an even larger ratio (2.79) but a much smaller absolute pass rate, illustrating why both quantities are needed. The mean independence prediction for PREFGATEGFN is 22.5%; the ratio of the two means is 2.46, distinct from the mean of target-wise ratios. These measurements motivate the matched library factorial rather than substituting for it.

F.3 ACCEPTABLE-REGION HYPERVOLUME

Docking magnitude $s = -d$ is transformed to $\mathrm{clip}((s-8)/(s_t^*-8), 0, 1)$, the physicochemical indicator is binary, and ADMET desirability u is transformed to $\mathrm{clip}((u-0.5)/0.5, 0, 1)$. Hypervolume is the volume dominated by these vectors in the unit cube. Thus a molecule below a reference threshold contributes no positive volume along that coordinate. The pinned docking anchors s_t^* are 10.09146, 12.84684, 10.74740, 10.78140, and 9.74443 kcal/mol for ADRB2, ALDH1, ESR agonist, ESR antagonist, and FEN1. They originate from pooled 99th-percentile docking magnitudes and are held fixed when comparing systems. Eight scrambled Sobol replicates of 65,536 points estimate each hypervolume; plot error bars summarize target variation, not the numerical integration error.

Mean PREFGATEGFN hypervolume changes from 0.582 to 0.234, while MOGFN-PC changes from 0.692 to 0.182. The corresponding cross-task ratios are 40.1% and 26.3%. The within-task comparison is the strongest interpretation: PREFGATEGFN has the highest measured mean coverage on the largest panel, but not the smallest.

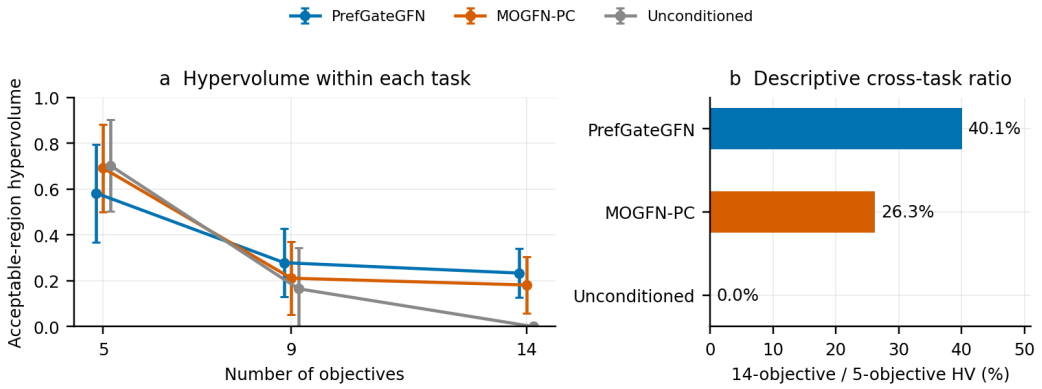

Figure 14: Acceptable-region hypervolume for 5, 9, and 14 objectives. Left: mean ± sample SD across five targets. Right: ratio of mean 14-objective to mean 5-objective hypervolume. These are differently dimensional regions with different criteria; the ratio summarizes this experiment series and is not a dimension-invariant retention measure.

G REUSING A FOURTEEN-OBJECTIVE POLICY FOR SMALLER PANELS

We compare the five- and nine-objective specialists with the fourteen-objective PREFGATEGFN checkpoint sampled using the corresponding smaller objective profile. No retraining is performed for reuse. The reuse evaluation has 256 draws per target, whereas the specialist records contain 1,000. To make the draw budgets comparable, we take 2,000 random size-256 subsets without replacement from each specialist record, deduplicate each subset, and average its metrics. These are subsampling replicates of one evaluation record, not new training seeds. The reuse side remains a single 256-draw record.

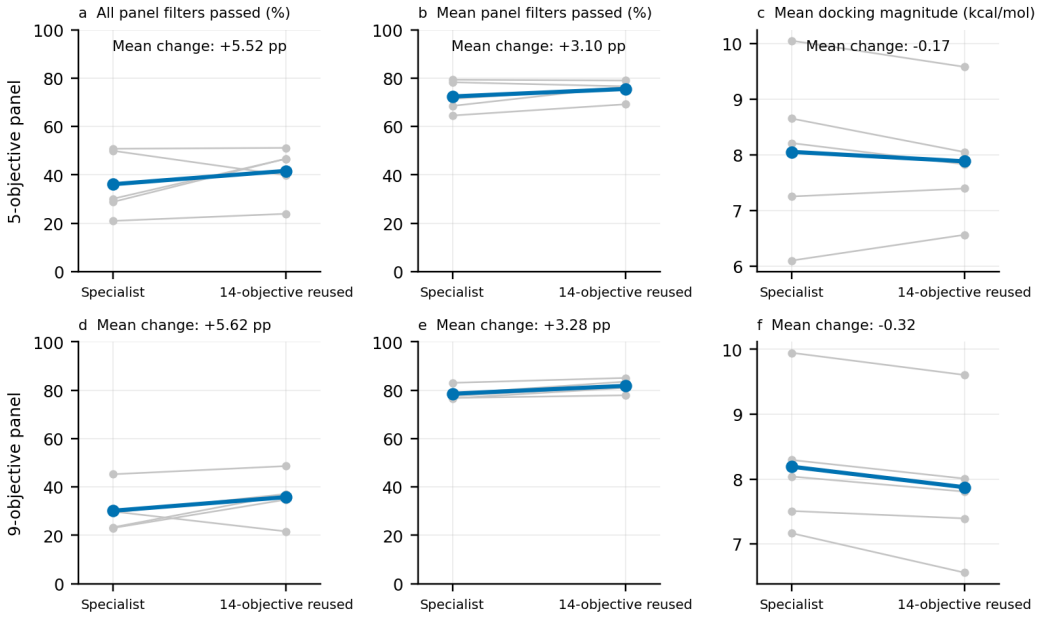

Figure 15: Objective-subset reuse with matched 256-draw budgets. Top: the five-objective profile. Bottom: the nine-objective profile. Grey lines pair individual targets; blue lines join equally weighted target means. Screening and mean filter-pass rates use unique valid molecules and only the requested panel's non-docking criteria. Docking magnitude averages returned finite negative scores, excluding docking failures. Specialist points average size-matched subsamples; reuse points use the recorded evaluation.

The reused model can maintain or improve the mean pass rates for these smaller panels while its mean docking magnitude is somewhat lower. Effects vary by target, so this is evidence of useful reuse, not a claim that a many-objective policy always dominates a specialist. In particular, a higher mean fraction of filters passed does not guarantee that more molecules pass every filter. Keeping those outcomes separate is important when the deployment decision requires their conjunction.

H REACTION PREFERENCE RESPONSE

H.1 JOINT DILI–BIOAVAILABILITY SCREENING COUNTS

The main-text joint-pass analysis (Figure 4) deduplicates valid molecules by canonical SMILES before applying $1 - p_{\text{DILI}} \geq 0.9$ and $p_{\text{bioavailability}} \geq 0.9$. Every count retains the original 1,000-draw budget as denominator; it is not conditional on a molecule being valid or on appearing on a frontier. The mean pairwise joint-pass counts are 29.8, 3.6, 2.2, and 1.4 for PREFGATEGFN, HN-GFN, concatenation, and MOGFN-PC, respectively. Their intersection with all thirteen non-docking screening filters gives 20.4, 0.8, 1.2, and 0.6; additionally requiring docking score ≤ -8 gives 12.6, 0.6, 0.4, and 0.2. These nested counts prevent a favorable pairwise projection from being mistaken for all-objective success. No new model fitting, docking, or ADMET inference is involved.

H.2 DIRECT REACTION PREFERENCE INTERVENTIONS

The sweep holds the checkpoint, sampling temperature, workflow policy, and random seed fixed within each target while varying a single objective's weight. The remaining mass is shared among other objectives. Each condition uses 256 draws, and the sweep covers ADRB2, ESR agonist, and FEN1. Outcomes are oriented toward desirability and standardized by their pooled within-target, within-panel, within-objective variation. We subtract the uniform-preference mean, average focused objectives within each target, and then average targets. Successful negative docking scores are converted to magnitudes; technical docking failures are excluded from that docking mean and are not interpreted as weak binding. Other invalid outcomes follow the stored evaluator's zero convention.

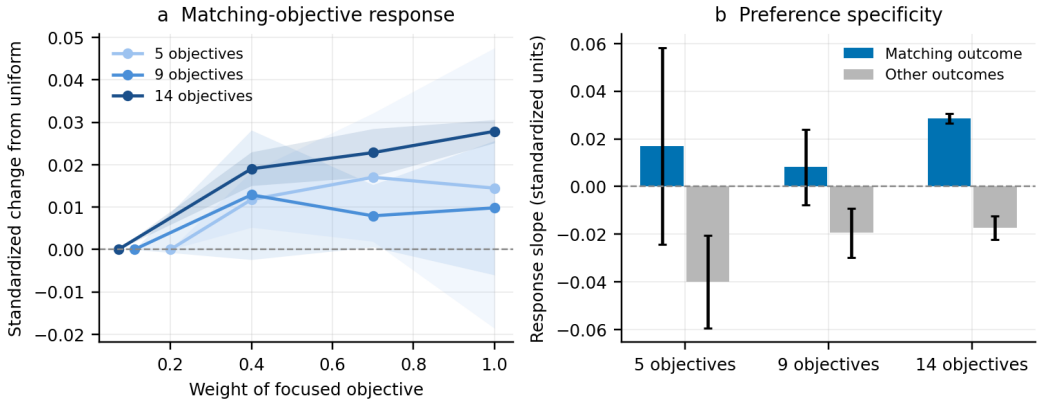

Figure 16: Direct preference response of PREFGATEGFN for the 5-, 9-, and 14-objective systems. Left: change in the requested outcome relative to uniform preferences. Right: slopes for the requested outcome and the average of other outcomes. Bands and bars show sample SD across three targets, not confidence intervals over independent model seeds. All panels use measured sampling outcomes, not internal gate values.

The 14-objective sweep has a positive mean matching response and a negative mean response in other outcomes, consistent with a realized trade-off. The standardized changes are small; the five- and nine-objective systems show larger between-target uncertainty. This single-method plot demonstrates that the conditioned system responds, not that it is the most responsive architecture. A separate paired 14-objective sweep over five targets gives mean matched response 0.055 ± 0.029 for PREFGATEGFN and 0.438 ± 0.218 for MOGFN-PC when the focused weight changes from 0.04 to 0.40. The unequal controller-update schedules described in Appendix F limit architectural interpretation, but the measured difference is retained: higher joint candidate yield does not imply greater preference sensitivity.

I ILLUSTRATIVE DOCKED CANDIDATES

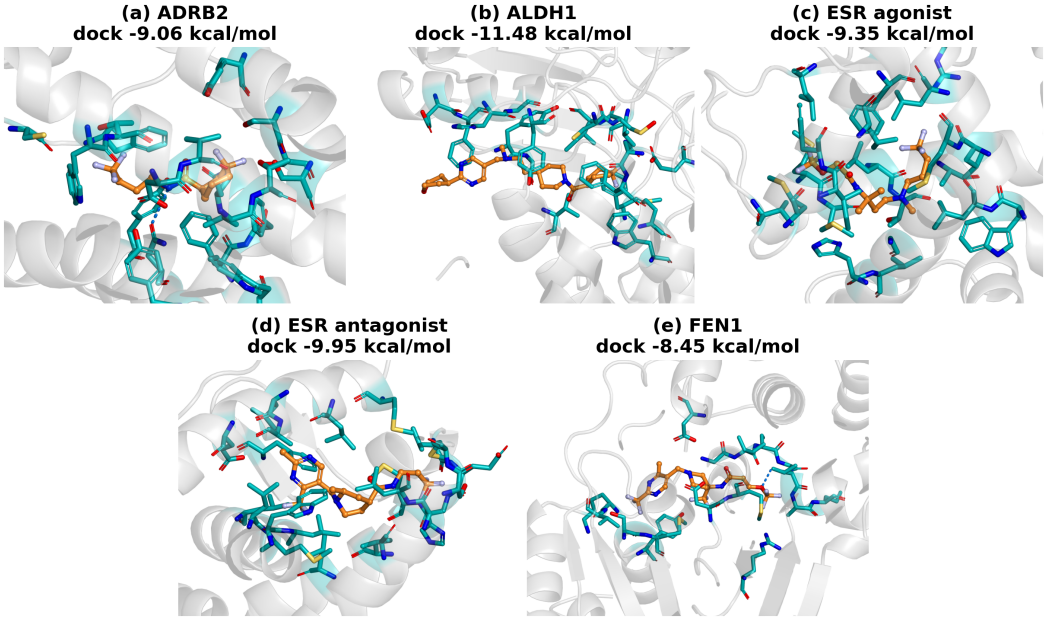

Figure 17: Illustrative PREFGATEGFN candidates in the five target structures. Ligand carbons are orange, nearby protein carbons teal, and the protein backbone light grey; heteroatoms retain element colors. Titles report the stored evaluation docking scores in kcal/mol. These are selected examples, not average poses, independent validation, or experimentally observed binding. No new docking calculation was run for this illustration.

The ADRB2 example has a one-step recorded synthesis route; the other four have two-step routes. Each shown pose is linked to a canonical molecule and stored evaluation entry in the candidate manifest. The ADRB2 pose originates from the nine-objective run; the other examples originate from fourteen-objective runs. Thus this is an illustration of route-linked candidates across the study, not a five-target panel sampled exclusively from one experiment. Evaluation energies differ from the stored pose-file energies by at most 0.072 kcal/mol for these examples; the titles consistently use evaluation energies.

J NINE-OBJECTIVE COMPARISON ACROSS GENERATION SYSTEMS

We also examine separately supplied FragGFN-SA (Bengio et al., 2021), SynFlowNet (Cretu et al., 2025), and RxnFlow (ICLR 2025) (Seo et al., 2025) runs. They target the same nine objective names as our intermediate panel: docking, physicochemical suitability, hERG, PAMPA, P-gp, BBB, Ames, DILI, and bioavailability. RxnFlow (ICLR 2025) denotes that separate implementation, not our unconditioned policy or MOGFN-PC. Table 6 compares all four systems over 1,000 recorded training iterations and approximately 64,000 draws per run. We use the existing fifteen-run PREFGATEGFN campaign at that duration, rather than final-checkpoint samples. For this table, we additionally apply their exclusions of charged molecules, triple bonds, and spiro atoms to our generated samples. Joint success requires the seven ADMET filters, the physicochemical/structural criteria, and docking ≤ -8 kcal/mol for every system.

Table 6: **Distinct qualifying molecules on the nine-objective task.** Counts are means per completed run over 1,000 recorded training iterations (63,865–64,000 draws/run), not percentages or final-checkpoint samples. All-7 ADMET counts distinct molecules passing all seven predicted endpoint filters. Joint hits additionally require common physicochemical/structural eligibility and docking ≤ -8 kcal/mol. Each qualifying recorded SMILES is counted once within a run; counts are then averaged equally across runs, not pooled across seeds or targets. Bold marks the highest mean, including ties, not statistical significance. Training recipes differ across systems.

Model	Runs	Unique all-7 ADMET/run	Unique joint hits/run
Training streams: 1,000 iterations; approximately 64,000 recorded draws/run			
FragGFN-SA	15	48.5	0.6
RxnFlow (ICLR 2025)	15	203.3	8.7
SynFlowNet	13	53.1	1.0
PrefGateGFN	15	**5802.5**	**329.7**

The docking requirement distinguishes predicted ADMET satisfaction from jointly qualifying candidates. For example, one molecule accounts for 62,995 of 64,000 draws in a FragGFN-SA ADRB2 run and passes all seven ADMET filters, but its docking score of -6.478 kcal/mol does not clear the common -8 cutoff. Reporting distinct qualifying molecules avoids equating repeated sampling of such a molecule with broad candidate discovery.

Both outcome columns in Table 6 report mean unique counts per run. The all-seven-ADMET column isolates simultaneous predicted ADMET satisfaction from the additional physicochemical/structural and docking requirements in the joint-hit column. We deduplicate qualifying recorded SMILES within each run before averaging; repeatedly generating the same qualifying molecule therefore contributes one, not many, candidates. The following protocol subsection retains draw-level endpoint rates and docking-threshold sensitivity as supplementary diagnostics.

Across its matched-duration training streams, PREFGATEGFN discovers 329.7 distinct jointly qualifying molecules per run, compared with 8.7 for RxnFlow (ICLR 2025), 1.0 for SynFlowNet, and 0.6 for FragGFN-SA. Before imposing the additional physicochemical/structural and docking requirements, it produces 5,802.5 distinct all-seven-ADMET hits per run, compared with 203.3, 53.1, and 48.5 for those systems, respectively. These counts measure the discovery of many distinct candidates that combine the predicted ADMET profile with physicochemical eligibility and stronger docking, not repeated draws of a small number of qualifying molecules. The following protocol subsection also retains the favorable-sign docking diagnostic for all systems, making threshold sensitivity explicit.

The comparison aligns the training-stream window and generated-draw budget. It evaluates complete generation systems, not architecture alone: exploration, losses, replay, gradient accumulation, and action spaces differ. Nine-objective results complement, rather than replace, the separate fourteen-objective evidence. The following subsection details completion checks, denominators, and protocol alignment.

J.1 Evaluation protocol and supplementary diagnostics

We analyze the supplied FragGFN-SA, RxnFlow (ICLR 2025), and SynFlowNet result directories. RxnFlow (ICLR 2025) is the independently supplied implementation associated with Seo et al. (2025); it is distinct from the systems evaluated in Figure 3. FragGFN-SA is the fragment-GFlowNet baseline with an additional synthetic-accessibility objective, rather than a separately published architecture (Bengio et al., 2021; Seo et al., 2025). The reward configuration activates docking, physicochemical suitability, hERG, PAMPA, P-gp, BBB, Ames, DILI, and bioavailability, matching the nine objective names in our intermediate panel. Its scalarization is a static product. The supplied physicochemical protocol additionally disallows charged molecules, triple bonds, and spiro atoms. Matching objective names therefore does not by itself establish identical eligibility or training conditions.

We reaggregate the recorded reward and component CSVs for each completed run rather than use a checkpoint's best molecule. There are fifteen completed FragGFN-SA runs, fifteen RxnFlow (ICLR 2025) runs, and thirteen SynFlowNet runs across the five targets. SynFlowNet ALDH1 seed 1 and FEN1 seed 1 stop after 60 and 62 recorded iterations, respectively, and are excluded from the completed-run summary. Completed streams contain 63,865–64,000 recorded draws across 1,000 recorded training iterations. Rates divide by the actual record count, retain duplicate draws, and treat invalid recorded draws as failures. Distinct success counts use distinct recorded SMILES within a run; they are not canonicalized across different runs.

Table 6 compares these streams with fifteen completed PrefGateGFN runs from the existing 1,000-iteration campaign: three seeds (0–2) per target, each with 64,000 recorded draws. All 1,000 batch files are checked for completeness and 64 rows per batch. The per-draw reward records contain the nine objective values and their ordered head names; we apply endpoint thresholds to these values, not to normalized rewards. The four physicochemical intervals (MW, TPSA, HBD, and LogP) are checked against the supplied protocol. RDKit additionally excludes formal charge, triple bonds, and spiro atoms from joint success; rejected draws remain in the denominator. The joint screen requires a recorded docking magnitude of at least 8, corresponding to a score of at most -8 kcal/mol; failed docking is excluded. The same cutoff is applied to all four systems. Table 6 reports mean unique counts for both all-seven ADMET satisfaction and the full joint screen. Qualifying recorded SMILES are deduplicated within each run before averaging equally over completed runs, not pooled across seeds or targets. Fractional table entries arise from averaging integer counts, not from fractional molecules. Supplementary draw-level rates retain duplicate draws and the original draw denominator. The training window and stricter structural screen explain why these nine-objective values are not interchangeable with checkpoint screening in Figure 3.

Table 7: Individual ADMET pass rates (%) over 1,000 training iterations on the nine-objective task, averaged equally over completed runs. These supplementary draw-level rates retain duplicates and count invalid draws as failures, unlike the unique counts in Table 6. Marginal endpoint rates do not determine simultaneous satisfaction. Docking and physicochemical/structural criteria are not imposed here. Bold marks the highest displayed mean for each endpoint.

Model	hERG	PAMPA	P-gp	BBB	Ames	DILI	Bioavail.
Training streams							
FragGFN-SA	**24.74**	25.55	25.13	25.34	25.15	25.23	24.92
RxnFlow (ICLR 2025)	6.05	26.56	7.83	20.29	25.04	5.85	25.63
SynFlowNet	20.59	68.13	**73.69**	68.05	67.97	22.69	75.11
PrefGateGFN	20.61	**83.08**	59.63	**81.65**	**80.62**	**62.05**	**79.68**

Individual endpoint passes use a 0.5 threshold: liability predictions (hERG, P-gp, Ames, and DILI) must be at most 0.5; positive-direction predictions (PAMPA, BBB, and bioavailability) must be at least 0.5. All-seven satisfaction is determined from the conjunction of these flags on each draw, not estimated by multiplying marginal rates. The resulting qualifying SMILES are deduplicated for Table 6. Structural eligibility is applied only to joint success, not to the ADMET-only columns.

Table 8 retains the less restrictive favorable-sign docking diagnostic alongside the common -8 kcal/mol screen. No runs or molecules are removed from rate denominators when changing the threshold. The favorable-sign criterion counts many repeated weakly docked molecules: two FragGFN-SA runs contribute 99.7% of its joint-success draws under that criterion. This explains why a high

Table 8: Docking-threshold sensitivity for the same nine-objective training streams. All seven ADMET and physicochemical/structural filters are held fixed. Rates are mean percentages of all recorded draws; distinct hits are mean within-run counts. The favorable-sign diagnostic is retained for transparency alongside the -8 kcal/mol screen used in Table 6.

	Docking < 0		Docking ≤ -8	
Model	Pass (%)	Distinct hits	Pass (%)	Distinct hits
FragGFN-SA	**8.440**	6.9	0.001	0.6
RxnFlow (ICLR 2025)	0.098	27.3	0.026	8.7
SynFlowNet	0.008	4.5	0.002	1.0
PrefGateGFN	1.399	**767.2**	**0.589**	**329.7**

draw-level success rate can coexist with a small number of distinct qualifying molecules. These computational screens do not establish binding affinity or ADMET properties experimentally.

These streams describe discovery during training, not generation from a fixed final policy. Matching 1,000 recorded iterations and approximately 64,000 draws removes the previous mismatch in evaluation windows, but does not equate optimization or compute budgets. For example, PREFGATEGFN uses gradient accumulation, replay, and a different exploration probability; the supplied systems also differ in losses, scalarization, and action spaces. This is a generation-system comparison, not an architecture-only ablation. The separate fourteen-objective checkpoint analysis remains the evidence for scaling beyond ten objectives. The completed-run audit, recomputed per-run metrics, and source paths accompany the analysis.