MIRRA: FAST, HIGH-FIDELITY MOLECULAR GENERATION WITH CHEMISTRY-INFORMED DISCRETE FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Fast, high-fidelity molecular generation requires accurate chemical choices and efficient sampling. Chemical building blocks reduce graph size, but make each node a choice among tens of thousands of related identities. We introduce MIRRA (Molecular Identity Representation, Refresh and Assembly), a discrete flow-matching generator that learns from chemical knowledge and evaluates identity probabilities only when needed, while keeping every identity in its vocabulary available. Chemical features share information across related blocks, and learned identity corrections distinguish them in molecular context. By deciding which positions will refresh before computing their probabilities, the sampler avoids unused predictions while preserving each sampling transition. Against the strongest baselines in each task, MIRRA reduces chemical and graph distribution errors on ZINC250k by 29–31% and 53–60%, respectively, with generation up to 8.6 times faster. Property targeting yields an estimated 7.8–20.8 times more distinct novel hits per second. On a reserved polymer test, MIRRA achieves 100% periodic validity with known chain topology, versus 57.0% for a polymer variational autoencoder. These results demonstrate fast, high-fidelity generation with the full chemical vocabulary available throughout sampling.

1 INTRODUCTION

Molecules reuse chemistry. Rings and functional groups recur across compounds, so selecting one building block can specify several atoms and bonds at once. This makes substructure-based generation attractive: it reduces the number of graph variables while preserving chemically meaningful units (Jin et al., 2018; Han et al., 2025; Degen et al., 2008; Lee et al., 2026). Compression of the graph, however, relocates the difficulty to the individual nodes. Instead of choosing among a few atom types, a model must predict an identity from tens of thousands of building blocks, each in the context of a different partially formed molecule.

Two requirements follow. First, the predicted probabilities must reflect both chemical similarity and individual difference. Two aromatic blocks may share useful structural information while differing in ring size, heteroatoms or attachment pattern, and these differences can determine which surrounding structures they fit. Second, evaluating those probabilities must remain affordable, since the vocabulary is scored at every position of every sampling step; a large vocabulary can otherwise erase the computational benefit of the smaller graph.

We ask: *can a generator retain detailed chemical choices and still be fast?* MIRRA (Molecular Identity Representation, Refresh and Assembly) answers by combining chemistry-informed prediction with refresh scheduling (Figure 1). Every identity in its vocabulary remains available throughout generation. The *representation* assigns each block a probability from two terms: fixed chemical features shared across related entries, and learned per-entry corrections that depend on molecular context. The *refresh* schedule fixes which positions will be resampled before any scoring occurs, so only those positions evaluate the full vocabulary while the complete graph still supplies context. *Assembly* converts the final block graph into a molecule by resolving compatible attachment sites from the terminal probabilities.

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

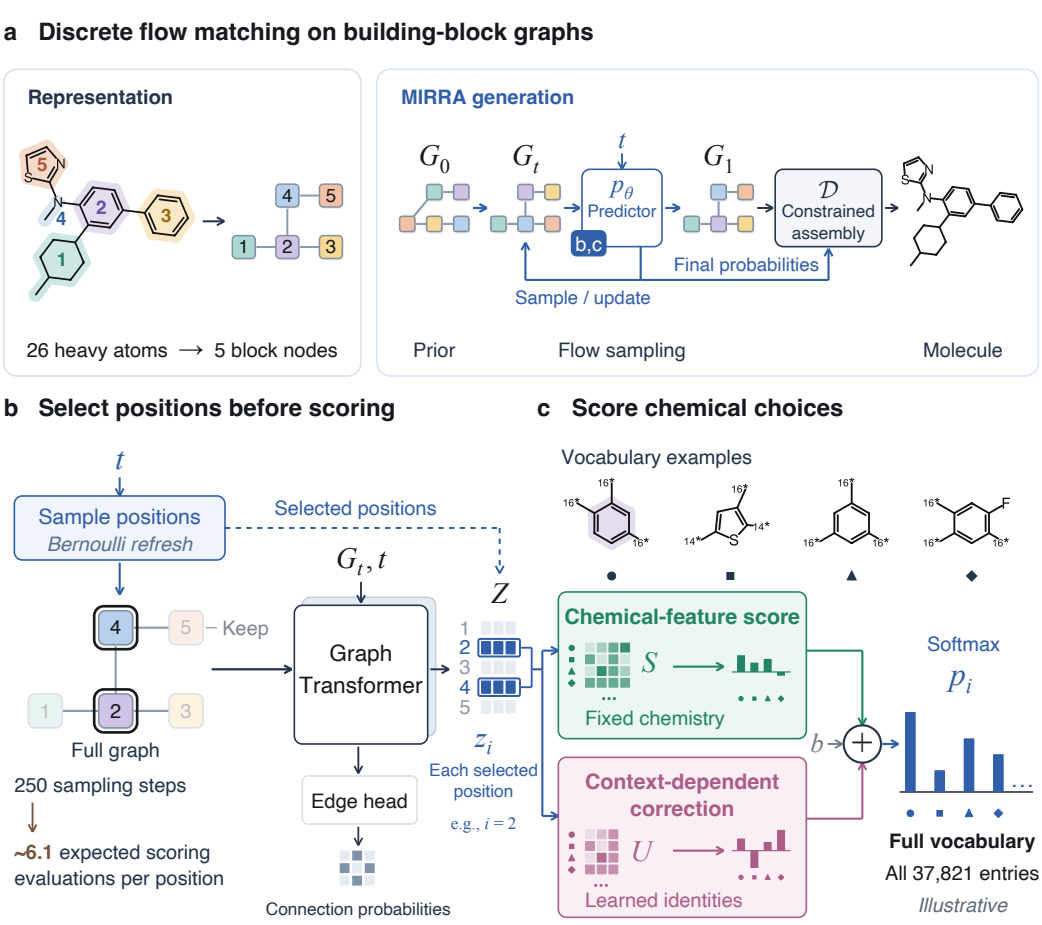

Figure 1: **MIRRA combines chemistry-informed prediction, refresh scheduling and constrained assembly.** (a) A 26-heavy-atom molecule becomes five building-block nodes. The predictor p_θ supplies node and connection probabilities for discrete flow matching, using refresh scheduling (b) and chemistry-informed identity scores (c). Constrained assembly $\mathcal{D}$ uses final identities and these probabilities to resolve compatible connections and attachment sites, with limited identity changes when needed. (b) Independent Bernoulli refresh decisions select positions for vocabulary scoring using the full graph as context. A selected identity can remain unchanged. At 250 steps, expected scores per position fall from 250 to 6.1; this is a computation count, not a runtime ratio. (c) Chemical-feature and identity-correction scores, plus a frequency offset, define probabilities over all 37,821 entries. The four chemical examples are vocabulary members, not a shortlist; intermediate states, matrices and probability bars are schematic.

We test MIRRA against these requirements:

- **Higher fidelity at lower cost.** MIRRA achieves the best chemical and graph fidelity in our ZINC250k benchmark. Compared with the strongest baseline, FragFM at 500 steps, MIRRA at 25 steps lowers Fréchet ChemNet Distance (FCD) by 28.7% and generates 8.6 times faster. Both models use the same training data, and matched timings include assembly.
- **An explanation supported by controls.** Scheduling preserves the finite-step kernel and gives a 3.0-fold speedup at fixed weights. Matched training controls establish the benefit of chemical alignment beyond capacity, and generation/assembly interventions locate the role of identity corrections.
- **Beyond the benchmark. Conditional design** increases estimated distinct novel target hits per second. On a reserved **polymer** test, MIRRA achieves **100% periodic validity** under known chain topology. The fixed design also transfers to MOSES.

2 MIRRA: REPRESENTATION, REFRESH AND ASSEMBLY

2.1 REPRESENTATION: CHEMICALLY STRUCTURED IDENTITIES

A molecule is represented by building-block (molecular fragment) identities $X = (x_1, \ldots, x_n)$ and a symmetric binary adjacency matrix E. Blocks expose typed attachment sites obtained from BRICS-derived decompositions (Degen et al., 2008), and their internal atoms and bonds are specified by identity, so the generator predicts only the block graph while distinct substitution and attachment patterns remain distinct choices. The benchmark vocabulary contains 37,821 entries; its finite size also bounds the chemical space this representation expresses.

Each entry v has a fixed vector s_v describing chemical structure and attachments. A graph transformer reads these features, the noisy edges and the time to produce a contextual state z_i for each node, and the model scores every entry as

$$\ell_i(v) = a\, s_v^\top g(z_i) + u_v^\top D z_i + b_v, \qquad p_i(v) = \mathrm{softmax}_v\, \ell_i(v). \tag{1}$$

The first term projects the contextual state into feature space through g with learned scale a, allowing related blocks to inform one another's learning. The second is the identity correction, a learned vector u_v for each entry read through a shared low-rank projection D. The fixed offset b_v encodes training frequency, and cross-entropy trains probabilities normalized over *every* vocabulary entry.

The two terms serve complementary purposes. For two entries with identical chemical features, the relative odds are fixed across all molecules without the correction; with it, their log-odds differ by $(u_v - u_w)^\top D z_i$ and depend on context, so ring similarity does not force two attachment patterns to be interchangeable. This is a statistical use of chemical structure rather than an energy or reaction model: it expands output expressivity without recovering distinctions absent from the input features (Appendix B.3).

Training uses the linear independent-coupling path of discrete flow matching (Gat et al., 2024): given clean identity x_1 and source prior p_0,

$$q_t(x_t \mid x_1) = t\, \delta_{x_1}(x_t) + (1 - t) p_0(x_t), \tag{2}$$

with an analogous symmetric binary corruption on edges. Individual identities remain explicit in the state throughout; chemical sharing changes their predictor, not the state. Appendix B specifies the features, dimensions and training recipe.

2.2 REFRESH: SCORING ONLY AT SCHEDULED POSITIONS

On the grid $t_k = k/T$, the finite-step node transition is

$$K_{i,k}(v \mid X, E) = (1 - \alpha_k)\mathbf{1}[v = x_i] + \alpha_k\, p_{\theta,i}(v \mid X, E, t_k), \qquad \alpha_k = \frac{1}{T - k}. \tag{3}$$

The transition is a mixture of retaining the current identity and drawing a new one from the model, and a conventional implementation evaluates $p_{\theta,i}$ before sampling from it. The mixture weight α_k, however, is fixed by the schedule and independent of the prediction, so the two operations may be reordered and the Bernoulli refresh drawn first. When the refresh is zero the identity is retained, and its vocabulary probabilities cannot affect the transition, so they need not be computed; when it is one, all 37,821 entries are scored and an identity is sampled, which may coincide with the current one. Edges are governed by analogous gates, all predictions use the incoming graph, and updates are simultaneous.

Because the gate is independent of the prediction, this ordering preserves Equation 3 in real arithmetic, and expected full-vocabulary projections per node decrease from T to $H_T = \sum_{j=1}^{T} 1/j$, approximately 6.1 instead of 250; instrumented runs reproduce this count (Appendix B.5). Early refreshes are sparse and the final step refreshes every node, while a graph in which nothing refreshes requires no forward pass. Within each active graph, the backbone and edge predictor remain dense; end-to-end speed is measured separately.

2.3 ASSEMBLY: FROM BLOCK GRAPH TO MOLECULE

The decoder uses the final node and edge posteriors to assemble the block graph into a molecule. A feasible wiring requires $d_i \le m(x_i)$, where d_i counts assigned connections and $m(x_i)$ the available

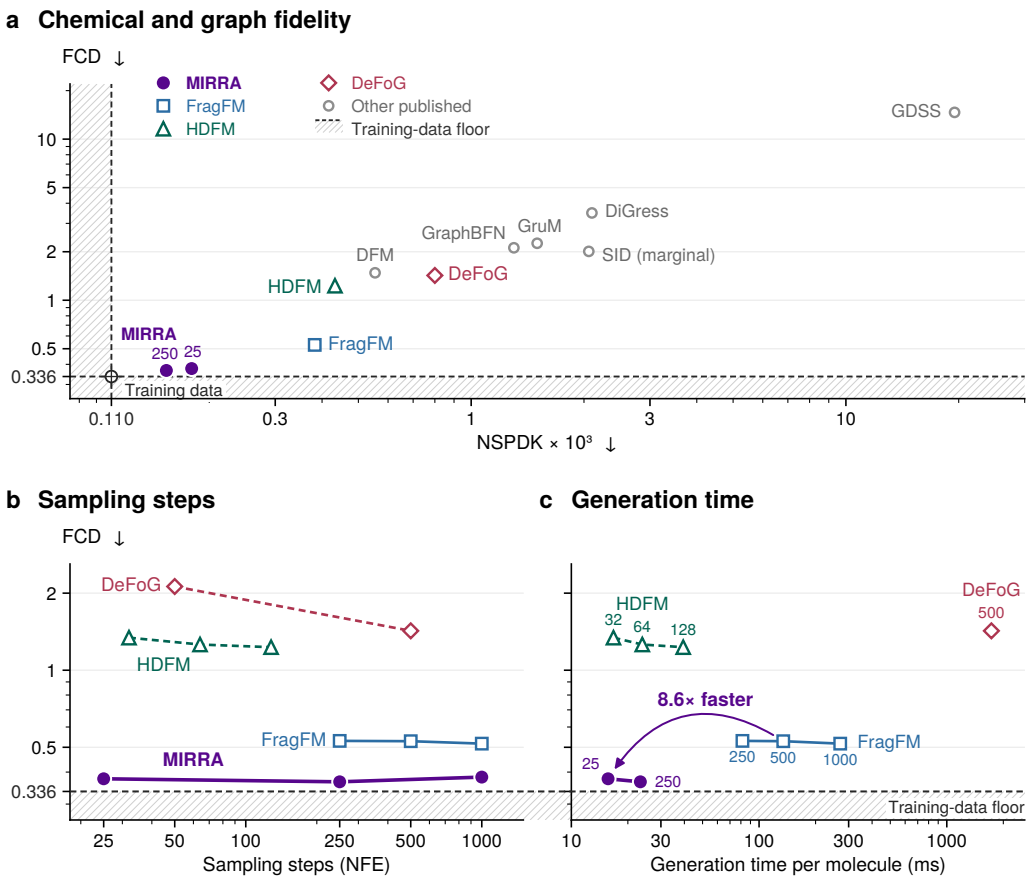

Figure 2: **Strong chemical and graph fidelity at low sampling cost.** (a) FCD and NSPDK discrepancies; (b) FCD versus steps (reported number of function evaluations, NFE, for HDFM); (c) FCD versus time per attempted molecule. Lower is better. Quality for MIRRA and FragFM uses 10,000 attempts; MIRRA250 is a three-draw median. MIRRA and FragFM suffixes denote sampling steps. Hatching marks the training-data floor, measured from a fixed training sample. The 8.6-fold arrow is the matched MIRRA25–FragFM500 timing comparison; their times are medians of four runs (three for FragFM250/1000). See Appendix A.

attachment sites; joined sites must also have compatible types. Guided by these scores and training-derived attachment preferences, it selects connections and sites, with limited identity repair when no valid wiring exists. The posteriors come from the final sampling transition, without another prediction. Assembly complements the learned trajectory (Section 4); failed assemblies remain in the attempted output denominator.

Relation to prior work. Fragment-based generation and functional-group coarse-graining precede MIRRA (Jin et al., 2018; Han et al., 2025; Lee et al., 2026), as do structured output predictors (Sun et al., 2017; Pappas & Henderson, 2019; Zhang et al., 2026). Prediction-independent event timing appears in first-hitting samplers for masked diffusion (Zheng et al., 2025b); sparse active-node and adaptive graph-flow sampling have precedents (Chen et al., 2023; Qin et al., 2025); and HDFM and HOG-Diff organize generation coarse-to-fine (Boget et al., 2026; Huang & Birdal, 2026). We bring prediction-independent scheduling to molecular fragment graphs whose identities can refresh repeatedly. Scheduling preserves the specified finite-step kernel while avoiding unneeded *full-vocabulary* projections rather than whole forward passes (Appendix I).

Table 1: **High fidelity at few sampling steps on ZINC250k.** Both MIRRA settings lead on FCD and NSPDK. **Bold**/underline mark best/second-best among displayed models, using unrounded values.

(a) Fidelity and validity: all baselines

Model (sampling steps)	FCD ↓	NSPDK $\times 10^3$ ↓	Valid (%) ↑
Training reference	0.336	0.110	n/a
GDSS (1000)	14.656	19.500	97.01
DiGress (1000)	3.482	2.100	94.99
GruM (1000)	2.257	1.500	98.65
GraphBFN (500)	2.116	1.300	99.22
SID, marginal (500)	2.010	2.060	<u>99.50</u>
DFM (128 NFE)	1.480	0.554	99.41
DeFoG (500)	1.425	0.800	99.22
HDFM (128 NFE)	1.230	0.433	99.30
FragFM (500)	0.528	0.382	99.14
MIRRA (25)	<u>0.377</u>	<u>0.180</u>	99.27
MIRRA (250)	**0.366**	**0.154**	**99.81**

(b) Uniqueness

Model (steps)	Unique (%) ↑
GDSS (1000)	99.64
DiGress (1000)	99.97
GruM (1000)	99.97
SID (500)	99.84
FragFM (500)	<u>99.98</u>
MIRRA (25)	**99.98**
MIRRA (250)	99.97

(c) Novelty

Model (steps)	Novel (%) ↑
GDSS (1000)	**100.00**
DiGress (1000)	<u>99.99</u>
GruM (1000)	99.98
SID (500)	99.97
FragFM (500)	99.23
MIRRA (25)	97.69
MIRRA (250)	97.61

Local rows: 10,000 attempts, 24,887 references; MIRRA250 is a three-draw per-metric median, other local rows one draw. Training reference is not ranked; validity is not applicable. Diversity panels use the same models with both metrics available. Sources and protocols: Appendix A.

3 High fidelity at low sampling cost

Evaluation and evidence units. We measure chemical distribution matching with FCD (Preuer et al., 2018), graph discrepancy with NSPDK MMD^2 (Jo et al., 2022), and validity, uniqueness and novelty separately. Table 1 and Figures 2–4 share one ZINC protocol: valid molecules from 10,000 attempts against all 24,887 references, failures not replaced. ZINC informed architecture and correction-rank selection, so MOSES and the reserved polymer test evaluate the fixed design (Appendix A).

3.1 The fidelity advantage survives few sampling steps

At 250 steps, MIRRA reaches FCD **0.3665** versus 0.5281 for FragFM at 500 steps, a **30.6%** reduction under the same local evaluation, and NSPDK$\times 10^3$ improves from 0.382 to 0.154. At 25 steps, FCD remains 28.7% below FragFM500 while matched generation is **8.6 times faster**, including assembly (Figure 2). These are complete generators trained on the same data, though capacities, representations and recipes differ.

The broader comparison places both MIRRA settings below every displayed baseline on FCD and NSPDK (Table 1), and MIRRA250 has the highest displayed validity at 99.81%, with 99.97% uniqueness and 97.61% novelty. These metrics support distinct claims: high validity alone does not establish distribution matching, and MIRRA does not maximize novelty. A separate study finds a larger 57.7% reduction under a 25,000-valid-output protocol, which uses a different population and does not supply the percentages reported here (Appendix C.4).

3.2 SCHEDULING, NOT RESTRICTION, MAKES THE VOCABULARY PRACTICAL

We run the benchmark checkpoint in three ways at 250 steps: score every entry at every position; score every entry only at scheduled positions; or score only a candidate subset at scheduled positions. The subset contains 1,536 distinct prior-sampled entries plus the current identities, renormalized over that set, and its runtime includes candidate construction. All three modes supply assembly with the same complete terminal scores.

Scheduling reduces median time from 131.6 to 43.7 ms per attempt, a **3.0-fold speedup**, with nearly identical observed FCD (0.3680 versus 0.3665). The candidate subset behaves differently: its FCD rises to 0.9469 and NSPDK$\times 10^3$ to 0.8780, and at 46.2 ms it is marginally slower than scoring all 37,821 entries under the same schedule (Figure 3), while saving 34.7% of peak allocated tensor memory. For this implementation and checkpoint, restricting the choices therefore yields memory savings rather than speed, at a substantial cost in fidelity; candidate-specific training and corrected proposals remain untested alternatives (Appendix D).

4 WHY THE COMBINATION WORKS

Table 2: **Correct chemical features improve both chemical and graph fidelity on ZINC250k.** Independently trained predictors share the input features, backbone, training budget and assembly rule, with comparable capacity. All rows use 250 sampling steps and the common final training update. Best values are bold; second-best values are underlined.

Output predictor	FCD ↓	NSPDK ↓ ($\times 10^3$)	Novel (%) ↑
Conventional learned predictor	0.1793	0.1223	**98.05**
Shuffled chemical features	0.1757	0.1547	97.55
Chemistry shuffled within matched groups	<u>0.1655</u>	<u>0.1188</u>	97.51
MIRRA: correctly assigned chemical features	**0.1561**	**0.1063**	<u>97.81</u>

Each row is one independently trained model at the common final update. Matched groups preserve attachment types and frequency bins. Broad-shuffle FCD gains replicate across three training pairs (**11–16%**). Distances use 25,000 valid outputs and the original reference convention, distinct from Table 1; novelty uses unique valid outputs.

Chemical alignment improves what the model learns. Chemical features could in principle help only by adding parameters. To separate chemical information from parameter count, we independently train predictors with correct or shuffled output-feature assignments, holding input features, individual labels, capacity and the training recipe fixed. Correct alignment improves FCD by **11–16% across three matched training pairs**. A narrower shuffle within attachment-type and frequency-bin groups also worsens fidelity, the more informative outcome: the benefit follows not from a structured output space in general, but from one that is chemically correct. At a common final update, MIRRA outperforms both shuffles and a conventional learned predictor on FCD and NSPDK (Table 2), while that predictor has slightly higher novelty (Appendix E).

Identity corrections guide the generation trajectory. The corrections could in principle act only at the end, where assembly reads the terminal probabilities. We therefore enable them independently during generation and during terminal assembly (Figure 4). Holding full assembly scores fixed, enabling generation corrections changes median FCD from 7.638 to 0.366, with paired reductions of 95.1–95.4%. Assembly-only corrections cannot recover this loss: they raise validity from 80.03% to 88.10% while FCD remains near 7.6. The model therefore uses contextual identity preferences during generation. These interventions show how the trained model uses its corrections; Table 2 compares independently trained predictors.

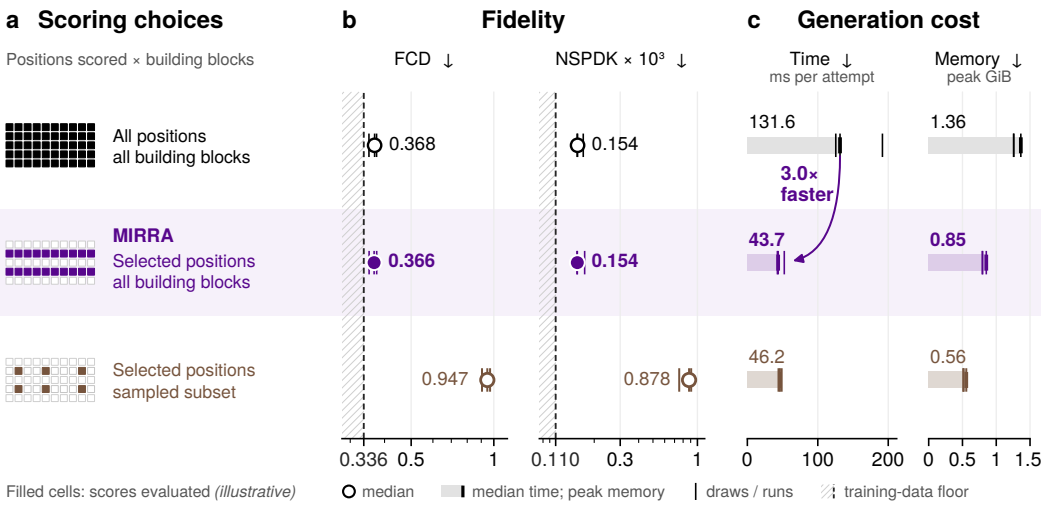

Figure 3: **Schedule the computation while preserving the choices.** (a) The positions and entries scored by three implementations of one checkpoint. (b) FCD and NSPDK: circles are three-draw medians; ticks show every draw. (c) Median time over three separate 1,024-attempt runs and peak allocated tensor memory, with individual runs shown. Scheduling is 3.0 times faster than dense scoring and slightly faster than the tested subset. Quality follows the 10,000-attempt protocol of Figure 2; timings use one GPU allocation and batch size 128. Candidate restriction has no separate training or proposal correction, and variation reflects sampling and timing repeats, not training-seed uncertainty.

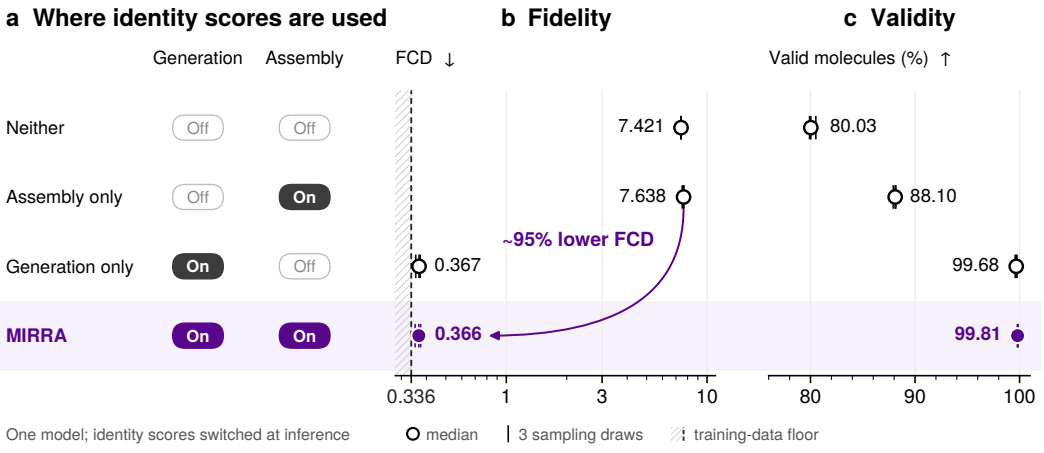

Figure 4: **Assembly cannot replace accurate identity choices during generation.** (a) Identity corrections are included or omitted during generation and assembly. (b) FCD on a logarithmic axis; (c) validity with an axis starting at 76%. Circles are medians and ticks show three sampling draws of the benchmark checkpoint. Assembly conditions within a generation arm share terminal states and edge probabilities. The full condition is exactly the MIRRA250 population in Figures 2–3; all use the same reference and evaluation. The roughly 95% FCD reduction compares full MIRRA with assembly-only correction. This fixed-weight intervention does not measure the effect of retraining without corrections.

Chemical validity does not imply the right distribution. Candidate restriction can preserve almost all valid outputs while changing molecular composition: in the three-model controls, a paired replay locates a molecular-size shift before assembly. An elementary argument explains why. Even if exact predictions preserve the target distribution, proposing one candidate from a nonuniform prior and renormalizing over it and the current identity can overrepresent common entries, and smaller

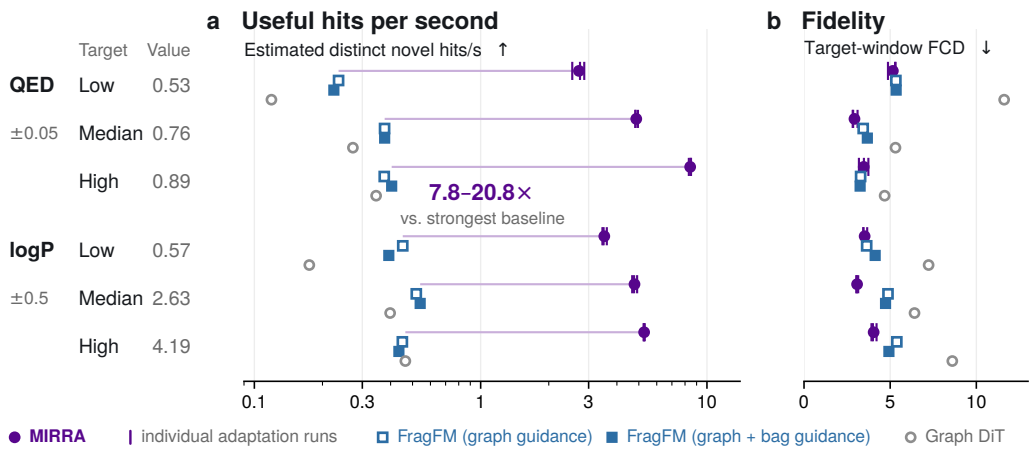

Figure 5: **Faster generation produces more useful property-targeted outputs.** (a) Estimated distinct novel hits per second; (b) target-window FCD, including valid misses. Markers average three adapter runs sharing one backbone; ticks show each run. Rates combine 4,096-attempt yields with three separate 256-attempt timings; FCD uses 1,000 valid outputs and 1,000 target references. Tolerances are ± 0.05 QED and ± 0.5 logP. Rates are estimates, not observed fixed-time searches.

time steps need not remove that bias. This concerns the specified choice rule rather than every candidate sampler (Appendix D.2).

5 What Fast, Faithful Generation Enables

Transfer beyond the development benchmark. On the official MOSES scaffold-test reference (Polykovskiy et al., 2020), the fixed design shows a mean FCD advantage over matched shuffled and conventional predictors, though the effect is smaller than on ZINC, varies with seed and step count, and comes with lower novelty (Appendix F).

More useful property-targeted outputs per unit time. For targeting the quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012) and logP, small property-dependent adapters modify transformer states while the benchmark backbone stays frozen. All methods use 500 steps and validation-selected guidance at the 10th, 50th and 90th training quantiles. A useful output is valid, lies within the target window and has a distinct identity absent from training; its estimated rate is

$$\widehat{r} = \frac{\text{distinct novel target hits/attempts}}{\text{separately measured seconds/attempt}}. \tag{4}$$

MIRRA produces an estimated **7.8–20.8 times** more distinct novel hits per second than the strongest tested baseline at each target (Figure 5); the range is 7.8–22.5 across both FragFM guidance settings and 11.4–24.4 against Graph DiT (Liu et al., 2024). The advantage does not arise from a higher hit rate: both FragFM settings reach every target more often per attempt, while MIRRA completes 21.3–25.3 times as many attempts per second as either of them, and 12.3–13.0 times as many as the fastest baseline at each target (Figure 8); throughput more than offsets the lower hit rate. MIRRA also has the lowest mean target-window FCD at five of six targets, with high QED favoring FragFM. These rates average three adapter runs on one backbone and combine separate yield and timing populations, so they are estimates rather than long-run discovery rates (Appendix G).

Valid periodic structures on a reserved polymer test. A split grouped by periodic identity reserves 9,949 test identities, with checkpoints and populations frozen before test evaluation. Every sampled MIRRA repeat unit is periodically valid, versus 57.03% for OMG-VAE (Kim et al., 2023) and 35.49% for Graph DiT (Table 3); periodic validity is stricter than parseability, requiring a connected repeat unit, two attachment ports and successful periodic canonicalization. At 250 steps the structural discrepancy is 0.021 in $\text{MMD}^2 \times 10^3$, with 98.30% novelty. MIRRA receives known linear chain topology while baselines generate full encodings, and two draws per checkpoint establish

Table 3: **Valid polymer repeat units with closer structural distributions.** Evaluation uses 9,949 reserved periodic identities. MIRRA has 100% periodic validity at both step counts. Best/second-best values are bold/underlined; ties share rank.

Model	Steps	Structural distance ↓ ($MMD^2 \times 10^3$)	Valid (%) ↑	Unique (%) ↑	Novel (%) ↑
OMG-VAE	Native	30.498 ± 0.458	<u>57.03</u>	97.32	**100.00**
Graph DiT	500	101.325 ± 2.048	35.49	**100.00**	**100.00**
MIRRA	500	<u>0.079 ± 0.030</u>	**100.00**	<u>99.96</u>	<u>98.57</u>
MIRRA	250	$\mathbf{0.021 \pm 0.015}$	**100.00**	99.91	98.30

Means over two 4,096-attempt draws of one checkpoint per method; distance uncertainty is sample SD. Scorable fractions are 56.10% for OMG-VAE, 35.49% for Graph DiT and 100% for MIRRA; distances condition on these outputs. Native denotes autoregressive decoding.

sampling rather than training-seed behavior (Appendix H). This demonstrates structural generation, not synthesis feasibility.

6 Discussion

MIRRA makes a large chemical vocabulary practical for fast, high-fidelity molecular generation. Chemical features let related building blocks share learning, while identity corrections preserve their differences in molecular context. Refresh scheduling makes these detailed choices affordable: it computes full-vocabulary probabilities only at positions where the transition will use them, preserving the finite-step kernel. Every identity remains available whenever a position refreshes.

The experiments connect this design to both better distribution matching and faster generation. Training controls establish the value of chemical alignment, and generation/assembly interventions show that contextual identity preferences guide the trajectory itself. In property targeting, higher throughput yields higher estimated rates of distinct novel hits. This connects generation speed to the practical yield of molecular search.

The evidence concerns a finite vocabulary. Published protocols differ, and novelty and per-attempt targeting do not uniformly favor MIRRA; broader transfer and candidate-specific training remain open. The central result is that detailed chemical choices can support both fidelity and speed: chemical knowledge improves their probabilities, and refresh scheduling removes the cost of predictions that generation does not use.

AI use statement

The authors developed the research ideas and made all experimental design decisions. AI assistants supported method implementation and code development, results analysis, manuscript and figure polishing, and literature retrieval and discovery. The authors take responsibility for the manuscript and all associated results and artifacts.

Reproducibility statement

Appendices A–H specify the data processing, architecture, training, sampling and evaluation procedures. SHA-256 cryptographic hashes identify the exact code, data and model files used. Replay checks reproduce generated molecules, evaluation scores and prepared datasets, and verify short training runs. We will release the complete source code under an open-source license following submission of the camera-ready manuscript.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.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. doi: 10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

Yoann Boget. Simple and critical iterative denoising: A recasting of discrete diffusion in graph generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4713–4736, 2025. URL https://proceedings.mlr.press/v267/boget25a.html.

Yoann Boget, Pablo Strasser, and Alexandros Kalousis. Hierarchical discrete flow matching for graph generation. arXiv preprint arXiv:2604.00236, 2026. URL https://arxiv.org/abs/2604.00236.

Xiaohui Chen, Jiaxing He, Xu Han, and Liping Liu. Efficient and degree-guided graph generation via discrete diffusion modeling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4585–4610, 2023. URL https://proceedings.mlr.press/v202/chen23k.html.

Fabrizio Costa and Kurt De Grave. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 255–262, 2010. URL https://icml.cc/Conferences/2010/papers/347.pdf.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: 10.1002/cmdc.200800178. URL https://doi.org/10.1002/cmdc.200800178.

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, pp. 133345–133385, 2024. doi: 10.52202/079017-4239. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f0d629a734b56a642701bba7bc8bb3ed-Paper-Conference.pdf.

Chuqin Geng, Li Zhang, Mark Zhang, Haolin Ye, Ziyu Zhao, and Xujie Si. Neural proposals, symbolic guarantees: Neuro-symbolic graph generation with hard constraints. arXiv preprint arXiv:2602.16954, 2026. URL https://arxiv.org/abs/2602.16954.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Ming Han, Ge Sun, Paul F. Nealey, and Juan J. de Pablo. Attention-based functional-group coarse-graining: a deep learning framework for molecular prediction and design. *npj Computational Materials*, 11:355, 2025. doi: 10.1038/s41524-025-01836-7. URL https://doi.org/10.1038/s41524-025-01836-7.

Yiming Huang and Tolga Birdal. HOG-Diff: Higher-order guided diffusion for graph generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2502.04308.

John J. Irwin and Brian K. Shoichet. ZINC – a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005. doi: 10.1021/ci049714+. URL https://doi.org/10.1021/ci049714+.

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*, pp. 2323–2332, 2018. URL https://proceedings.mlr.press/v80/jin18a.html.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4839–4848, 2020. URL https://proceedings.mlr.press/v119/jin20a.html.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10362–10383, 2022. URL https://proceedings.mlr.press/v162/jo22a.html.

Jaehyeong Jo, Dongki Kim, and Sung Ju Hwang. Graph generation with diffusion mixture. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22371–22405, 2024. URL https://proceedings.mlr.press/v235/jo24b.html.

Alexia Jolicoeur-Martineau, Aristide Baratin, Kisoo Kwon, Boris Knyazev, and Yan Zhang. Any-property-conditional molecule generation with self-criticism using spanning trees. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=QGZd5Bfb1L.

Mohamed Amine Ketata, Nicholas Gao, Johanna Sommer, Tom Wollschläger, and Stephan Günnemann. Lift your molecules: Molecular graph generation in latent Euclidean space. In *International Conference on Learning Representations*, pp. 16890–16921, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2af641762dc02035c31a9314b2d090b6-Paper-Conference.pdf.

Seonghwan Kim, Charles M. Schroeder, and Nicholas E. Jackson. Open macromolecular genome: Generative design of synthetically accessible polymers. *ACS Polymers Au*, 3(4):318–330, 2023. doi: 10.1021/acspolymersau.3c00003. URL https://doi.org/10.1021/acspolymersau.3c00003.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947. URL https://doi.org/10.1088/2632-2153/aba947.

Joongwon Lee, Seonghwan Kim, Seokhyun Moon, Hyunwoo Kim, and Woo Youn Kim. FragFM: Hierarchical framework for efficient molecule generation via fragment-level discrete flow matching. In *International Conference on Learning Representations*, pp. 90722–90777, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/92f556b408cfac2943fb228207487967-Abstract-Conference.html.

Tzyy-Shyang Lin, Connor W. Coley, Hidenobu Mochigase, Haley K. Beech, Wencong Wang, Zi Wang, Eliot Woods, Stephen L. Craig, Jeremiah A. Johnson, Julia A. Kalow, Klavs F. Jensen, and Bradley D. Olsen. BigSMILES: A structurally-based line notation for describing macromolecules. *ACS Central Science*, 5(9):1523–1531, 2019. doi: 10.1021/acscentsci.9b00476. URL https://doi.org/10.1021/acscentsci.9b00476.

Gang Liu, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph diffusion transformers for multi-conditional molecular generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8065–8092, 2024. doi: 10.52202/079017-0260. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0f6931a9e339a012a9909306d7c758b4-Paper-Conference.pdf.

Ethan Ma, Zihan Wang, Chris Siu Yeung Chow, Xinguo Feng, Qingqing Li, Rui Jiang, Naipeng Dong, and Guangdong Bai. Embedded graph flows for categorical graph generation. arXiv preprint arXiv:2609.05328, 2026. URL https://arxiv.org/abs/2609.05328.

Farshad Noravesh, Reza Haffari, Layki Soon, and Arghya Pal. VQ-SAD: Vector quantized structure aware diffusion for molecule generation. arXiv preprint arXiv:2605.00354, 2026. URL https://arxiv.org/abs/2605.00354.

Nikolaos Pappas and James Henderson. Deep residual output layers for neural language generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5000–5011, 2019. URL https://proceedings.mlr.press/v97/pappas19a.html.

Moritz Piening and Christian Wald. Gromov–Monge flow matching for equivariant graph generation. arXiv preprint arXiv:2608.26961, 2026. URL https://arxiv.org/abs/2608.26961.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Simon Johansson, Hongming Chen, Sergey Nikolenko, Alán Aspuru-Guzik, and Alex Zhavoronkov. Molecular sets (MOSES): A benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020. doi: 10.3389/fphar.2020.565644. URL https://doi.org/10.3389/fphar.2020.565644.

Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet ChemNet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018. doi: 10.1021/acs.jcim.8b00234. URL https://doi.org/10.1021/acs.jcim.8b00234.

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*, pp. 50269–50326, 2025. URL https://proceedings.mlr.press/v267/qin25d.html.

Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail İlkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. Categorical flow maps. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61953.

Hyunjin Seo, Taewon Kim, Sihyun Yu, and Sungsoo Ahn. Learning flexible forward trajectories for masked molecular diffusion. In *International Conference on Learning Representations*, pp. 95424–95448, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9a17c1eb808cf012065e9db47b7ca80d-Paper-Conference.pdf.

Yuxuan Song, Juntong Shi, Jingjing Gong, Minkai Xu, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Smooth interpolation for improved discrete graph generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 56363–56388, 2025. URL https://proceedings.mlr.press/v267/song25f.html.

Xu Sun, Bingzhen Wei, Xuancheng Ren, and Shuming Ma. Label embedding network: Learning label representation for soft training of deep networks. arXiv preprint arXiv:1710.10393, 2017. URL https://arxiv.org/abs/1710.10393.

Yogesh Verma, Samuel Kaski, Markus Heinonen, and Vikas Garg. Modular flows: Differential molecular generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12409–12421, 2022. doi: 10.52202/068431-0901. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/509f7977030f3550300f541ec228c3fc-Paper-Conference.pdf.

Clément Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. DiGress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.14734.

Yida Xiong, Jiameng Chen, Xiuwen Gong, Jia Wu, Shirui Pan, and Wenbin Hu. Variational Bayesian flow network for graph generation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64460.

Gervasio Zaldivar, Yuan Tian, Ge Sun, Chryssa Kappatou, Philipp Eiden, Niklas B. Wulkow, Volker Settels, and Juan J. de Pablo. G^2RINS: A generative string-and-graph polymer representation to assist computational materials discovery. ChemRxiv preprint, version 1, 2026. URL https://chemrxiv.org/doi/full/10.26434/chemrxiv.15007504/v1.

Shuai Zhang, Weibo Xu, Jiahao Nie, and Kecheng Huang. LGSA: Label geometry structuring and aligning for hierarchical text classification. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28654–28665, 2026. doi: 10.18653/v1/2026.acl-long.1322. URL https://aclanthology.org/2026.acl-long.1322/.

Jiachi Zhao, Zehong Wang, Yamei Liao, Chuxu Zhang, and Yanfang Ye. Controllable graph generation with diffusion models via inference-time tree search guidance. In *Proceedings of the ACM Web Conference 2026*, pp. 1195–1205, 2026. doi: 10.1145/3774904.3792489. URL https://doi.org/10.1145/3774904.3792489.

Haozhuo Zheng, Cheng Wang, and Yang Liu. Graph VQ-Transformer (GVT): Fast and accurate molecular generation via high-fidelity discrete latents. arXiv preprint arXiv:2512.02667, 2025a. URL https://arxiv.org/abs/2512.02667.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations*, pp. 63186–63227, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9e3b203e72c4e058de26d02a92a81844-Paper-Conference.pdf.

APPENDIX CONTENTS

The supplementary material follows the main argument: evaluation, the model, the benchmark, the controlled explanations, and the applications. Section titles and page numbers below are clickable.

Section and contents	Page
A Evaluation protocols and evidence map	15
Populations, metrics, timing scopes and the displays they support	
B Model, training and exact scheduling	16
Representation, architecture, output algebra, sampler and assembly	
C Benchmark results and comparison protocols	18
Step counts, sampling variability, runtime and historical context	
D Full-vocabulary execution and candidate restriction	22
Matched cost and fidelity, composition shift and oracle calculation	
E Chemical alignment and identity corrections	23
Retrained controls, trajectory/assembly intervention and replications	
F Transfer to MOSES	26
Three-seed matched heads, checkpoint sensitivity and novelty	
G Property-conditioned generation	27
Adapters, guidance, six-target results, controls and useful rates	
H Polymer generation	31
Reserved split, baseline adaptations, metric and internal controls	
I Related work	34
Method distinctions and connections to prior work	

Reading the evidence. Table 1 and Figures 2–4 share the current ZINC benchmark checkpoint and scoring convention. Table 2 instead tests independently trained predictors under a separate protocol. Applications have their own reference populations. Sampling draws, independent training runs and timing repetitions are identified separately throughout.

A Evaluation protocols and evidence map

A.1 Current ZINC displays

ZINC250k is a molecular-generation benchmark derived from the ZINC database (Irwin & Shoichet, 2005). Table 1 and Figures 2–4 score valid molecules among the first 10,000 attempted outputs against all 24,887 reference molecules, with reference stereochemistry retained. Failures are not replaced; duplicate valid molecules remain in FCD and NSPDK. Validity divides valid outputs by attempts, uniqueness divides unique identities by valid outputs, and novelty divides unseen unique identities by unique identities against the common training reference. FCD measures ChemNet feature distributions; NSPDK is the squared maximum mean discrepancy of labeled local graph features, reported as $\mathrm{MMD}^2 \times 10^3$ (Preuer et al., 2018; Costa & De Grave, 2010; Jo et al., 2022).

MIRRA250 is the per-metric median of three draws of one checkpoint, with 9,981, 9,985 and 9,980 valid outputs. Draws are not pooled, and different metrics can select different median draws. FCD is 0.3665 (range 0.3514–0.3745); NSPDK is 0.1537 (0.1530–0.1723). These medians are shared across all current displays. MIRRA25 and FragFM500 use single draws with 9,927 and 9,914 valid outputs and FCD 0.3766 and 0.5281, giving reductions of 28.68% and 30.60% for MIRRA25 and MIRRA250, respectively.

A fixed 10,000-molecule training draw defines the training-data floor: FCD 0.3361 and NSPDK 0.1096. ZINC informed architecture and correction-rank selection; MOSES and the reserved polymer reference assess the fixed design.

A.2 Populations behind the other claims

Independently trained predictors (Table 2). The matched heads use final exponential moving average (EMA) weights, 250 steps, prior assembly, and the first 25,000 valid outputs from 25,600 attempts, with reference stereochemistry removed (Appendix E). Their FCD 0.1561 is not another estimate of the benchmark checkpoint's 0.3665. The earlier 57.7% FragFM comparison also uses the 25,000-valid convention; its separate population is documented in Appendix C.4.

Property targeting (Figure 5). Three property-adaptation runs share one frozen benchmark backbone; each baseline regime is one trained model. Every cell uses 4,096 attempts. Target-window FCD scores the first 1,000 valid outputs, including misses and duplicates, against 1,000 held-out molecules inside the requested window. Distinct novel hits count each valid, in-window, non-training identity once, ignoring stereochemistry. Each adaptation's yield is divided by its separately measured time per attempt before rates are averaged. The primary grid, two-seed shuffled-label controls and timing protocol are in Appendix G.

Polymers (Table 3). The reference contains 9,949 reserved periodic identities; each setting uses two 4,096-attempt draws of one checkpoint. Structural distances condition on scorable outputs; validity retains all attempts. MIRRA receives a known linear chain topology, while the baselines generate complete repeat-unit encodings. Appendix H specifies the split, periodic canonicalization, metric and separate validation controls.

A.3 Timing scopes and uncertainty

Between-generator timings in Figure 2 use batch 256 and four 4,096-attempt repetitions per primary comparison; the FragFM250/1000 extensions use three repetitions. Figure 3 instead uses batch 128 and three 1,024-attempt repetitions per arm. Both include initialization, generation, transfer, assembly and strict acceptance, excluding loading, warmup, writing and distribution scoring. Thus 23.32 ms for MIRRA250 in the first study and 43.74 ms in the second refer to different batching and execution conditions.

Conditional timing uses batch 128 and three 256-attempt repetitions per cell, including generation, decoding and the common property, identity and hit checks. Methods within each matched study use the same physical GPU and balanced execution order, but retain their pinned software stacks (Appendix B.2). Published timings retain their own hardware and scope. Timing repetitions do not

estimate quality uncertainty; sampling spread describes fixed checkpoints, and training variation is reported only for independently trained models.

B MODEL, TRAINING AND EXACT SCHEDULING

B.1 REPRESENTATION AND ARCHITECTURE

A molecule is represented by fragment identities $X = (x_1, \ldots, x_n)$ and a symmetric binary adjacency matrix E. BRICS-derived decompositions expose typed attachment sites (Degen et al., 2008); identities specify internal atoms and bonds, while assembly resolves inter-fragment connections. The current benchmark vocabulary has 37,821 entries. Historical models and matched ZINC head controls use a separate 38,951-entry vocabulary.

Each identity has a fixed 527-dimensional structural and attachment vector s_v. The graph transformer reads s_{x_i}, noisy edges, time and engineered graph features. It has 12 layers, eight attention heads, and node/edge/global widths 384/96/64. Its contextual node state z_i is scored as

$$\ell_i(v) = a\, s_v^\top g(z_i) + u_v^\top D z_i + b_v, \qquad a = \exp(\tau), \quad p_i = \operatorname{softmax} \ell_i. \tag{5}$$

The fixed offset b_v is log training frequency floored at 10^{-12}. The correction UDz_i has rank at most 256; U starts at zero. Both branches score every vocabulary entry. In the 38,951-entry vocabulary, 920 entries share a feature row with another identity: categorical labels are retained in the state, but their feature encoding is not injective.

B.2 TRAINING AND COMPUTE ENVIRONMENT

The linear independent-coupling path is

$$q_t(x_t \mid x_1) = t\, \delta_{x_1}(x_t) + (1-t) p_0(x_t). \tag{6}$$

Binary edges undergo analogous corruption, sampled on the upper triangle and mirrored. The objective is mean node cross-entropy plus five times mean edge cross-entropy, excluding padding and diagonal edges. Source priors are fixed within each comparison. Matched ZINC controls use 213,340 optimizer rows, 11,228 internal-validation rows, and an 8,192-row checkpoint-selection subset. They train for 150 epochs at batch 128 (249,900 updates), using AdamW with learning rate 2×10^{-4}, one-epoch warmup, weight decay 10^{-12} and EMA decay 0.9995. Appendix E distinguishes validation selection from the common-final comparison.

The common-cohort benchmark fixes 208,302 optimizer-training and 10,926 validation source rows for both MIRRA and FragFM. Novelty uses 207,114 unique canonical training identities. MIRRA has 27.2M parameters and trains for 150 epochs (244,050 updates); FragFM uses 100,000 autoencoder and 100,000 flow updates, with 21.7M flow parameters and a frozen 6.6M-parameter autoencoder. Data and evaluation are controlled, while architectures, tokenizations and training recipes differ.

Three historical checkpoints use separate 213,340/11,228 internal splits, ten-epoch warmup, epoch-wise cosine decay and training times $t = u^{0.7}$ for uniform u. They retain final-epoch EMA. Their initial log-frequency offset uses the 224,568-row standard training corpus; corruption priors use each optimizer subset. These checkpoints are distinct from the current benchmark and matched-head models.

Training and sampling use RTX6000 and H100 GPUs. MIRRA's stack is PyTorch 2.10.0, CUDA 12.8 and RDKit 2025.03.3; FragFM retains PyTorch 2.1.0, CUDA 11.8 and RDKit 2023.09.2. Matched unconditional timing uses one RTX6000 per comparison, four CPU threads, synchronized boundaries and balanced order. Cross-model throughput therefore includes software differences. Peak GPU memory means PyTorch allocated tensor memory after warmup, including persistent model tensors, excluding CPU memory and allocations outside that allocator.

B.3 WHAT CHEMICAL SHARING AND IDENTITY CORRECTION EXPRESS

Write $\ell = aSh + Uc + b$, with $h = g(z)$, $c = Dz$ and $d = p - e_y$ for target y. Cross-entropy gives

$$\nabla_h L = aS^\top d, \qquad \nabla_U L = dc^\top, \qquad \nabla_D L = (U^\top d) z^\top. \tag{7}$$

For a feature-coordinate step $h' = h - \eta \nabla_h L$ holding other factors fixed,

$$\ell'_v - \ell_v = \eta a^2 \left(s_v^\top s_y - \sum_w p_w s_v^\top s_w \right). \tag{8}$$

Thus $SS^\top$ couples label errors through chemical similarity. Permuting its feature rows preserves the spectrum but changes which identities share those directions. Trained alignment controls test the assignment; this partial coordinate update is not the full AdamW update or a causal decomposition of FCD.

If $s_v = s_w$, the structural head alone fixes their log-odds at $b_v - b_w$. The correction adds $(u_v - u_w)^\top Dz$, allowing context-dependent odds. More generally, if P projects onto $\mathrm{col}([\mathbf{1}, S])$, then $(I-P)UDz$ contains score directions the structural branch cannot express, even allowing a common logit shift. Appendix E.3 tests their usefulness. A label with no positive optimizer examples still receives residual gradient $p_v c$ through softmax competition.

The head for the 38,951-entry vocabulary has 10.4M trainable parameters versus 15.1M for a hypothetical $384 \to 384 \to 38{,}951$ dense MLP, but requires 2.05 times as many matrix multiply–accumulates per scored node. Parameter sharing alone does not establish a runtime advantage; the matched execution experiment does.

B.4 Sampler and kernel preservation

Sampling initializes graph sizes, identities and edges from fixed source laws. On $t_k = k/T$, $k = 0, \ldots, T-1$, independently draw node and upper-triangular edge gates with probability $\alpha_k = 1/(T-k)$. For each graph with any active coordinate, evaluate the backbone on its incoming state. Evaluate full-vocabulary scores only for active nodes; sample their identities and the active binary edges, then commit all updates simultaneously and mirror the edges. An active node can resample its current identity. Graphs without active coordinates require no forward pass.

This realizes exactly

$$K_{i,k}(v \mid X, E) = (1 - \alpha_k)\mathbf{1}[v = x_i] + \alpha_k p_{\theta,i}(v \mid X, E, t_k) \tag{9}$$

in real arithmetic: a zero gate retains the identity regardless of its unused prediction; a one gate draws from the complete posterior. Gate independence permits the computation order to change. Graphs can also be batched at their next active times because inference has no interaction between batch members.

For n nodes and $m = n(n-1)/2$ undirected edges, the expected numbers of node projections and graph evaluations are respectively

$$nH_T = n \sum_{j=1}^{T} \frac{1}{j}, \qquad \sum_{k=0}^{T-1} \left[1 - (1 - \alpha_k)^{n+m} \right]. \tag{10}$$

The first replaces nT dense projections. The backbone and edge head remain dense within active graphs; batching and assembly still cost time. Rebatching changes floating-point arithmetic, so kernel equality does not imply bitwise-identical trajectories. Dense/scheduled checks locate rare output differences at categorical CDF boundary crossings. Their quality is measured separately in Appendix D.

B.5 Measured refresh counts and final assembly

Figure 6 checks the benchmark checkpoint at 250 and 25 steps. Mean per-node counts agree with H_T within 0.24% and 0.10%, using 1,024 and 128 attempted molecules. This verifies the implementation of the schedule; the 3.0-fold runtime gain comes from the separate timing study.

At $k = T-1$, every node refreshes. Complete node and edge posteriors on that final incoming graph are retained for assembly. Assembly checks attachment counts and compatible typed ports, resolves connections and attachment sites, and attempts limited identity repair when no valid wiring exists. The prior-based policy uses training-derived attachment preferences; failures remain in the attempt denominator. Appendix E.2 separates correction during the trajectory from correction in terminal scores.

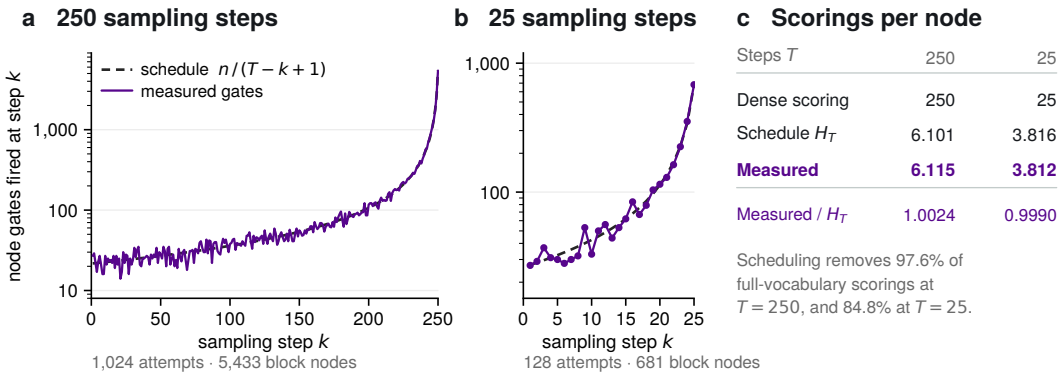

Figure 6: **Measured refresh counts follow the $1/(T-k+1)$ schedule, so the expected cost per node is H_T rather than T.** (a) At $T = 250$, node gates fired at each step against the schedule's expectation $n/(T-k+1)$ for n block nodes and one-based step index $k = 1, \ldots, T$, on a logarithmic axis. Early steps refresh few positions and the final step refreshes every one. (b) The same at $T = 25$. (c) Full-vocabulary scorings per node. Dense scoring evaluates the vocabulary at every step; the schedule predicts the harmonic number H_T; the measured counts are the instrumented totals divided by the number of block nodes. Counts come from instrumented sampling runs of the benchmark checkpoint (1,024 attempts at $T = 250$, 128 at $T = 25$) and record the Bernoulli gates that fired, which is exactly the set of positions at which the vocabulary is scored. The agreement is a check that the implementation draws the gate before the prediction as Equation 3 permits; it is not itself evidence of the end-to-end speedup, which Table 7 measures, because the backbone and edge predictor remain dense within an active graph.

C Benchmark results and comparison protocols

C.1 Current operating points and sampling variability

Table 4 collects the local operating points behind Figure 2. MIRRA has lower FCD and NSPDK at both shared step counts, 250 and 1000; the separate matched timings establish its practical advantage at 25 and 250 steps. FragFM retains higher novelty. All rows use the current 10,000-attempt protocol of Appendix A.

Eight sampling draws of the benchmark checkpoint at 250 steps give mean FCD 0.3659 and sample SD 0.0068 (range 0.3514–0.3745), compared with FragFM500's single-draw 0.5281. Their FCDs are 0.3514, 0.3665, 0.3745, 0.3693, 0.3674, 0.3649, 0.3704 and 0.3629. Mean NSPDK is 0.1610 ± 0.0073; mean validity, uniqueness and novelty are 99.82%, 99.94% and 97.54%. The first three draws supply the displayed per-metric medians. This spread measures sampling from one checkpoint, not uncertainty over training runs. The current checkpoint's step sweep is not monotone: FCD improves from 25 to 250 steps, then rises slightly at 1000.

C.2 Matched generation time and memory

Table 5 retains every timing repetition for the primary local comparisons and the FragFM extensions. Timers include all attempts and one assembly policy, with the scope in Appendix A. The matched 25-versus-500 experiment gives median times 15.69 and 134.53 ms, a factor of 8.58. The separate 250-versus-500 experiment gives 23.32 and 132.43 ms, a factor of 5.68. The first run of the former is slower in both arms and remains included. Quality populations are separate from timing runs.

In the 250-versus-500 study, median peak PyTorch allocation is 1.47 versus 2.73 GiB, 46.0% lower for MIRRA. This counter includes persistent model tensors but excludes loading and warmup peaks, CPU memory and allocations outside PyTorch. It does not measure minimum required device memory.

Table 4: **Complete ZINC250k sampling and evaluation grid.** All settings retained for both generators trained on the same data. Bold marks the best value in each column, using unrounded values.

Setting	Valid n	FCD ↓	NSPDK ↓ ($\times 10^3$)	Valid (%)	Unique (%)	Novel (%)	ms/attempt ↓
Common training data, one evaluation protocol							
MIRRA, 25 steps	9,927	0.3766	0.1795	99.27	**99.98**	97.69	**15.7**
MIRRA, 250 steps	9,980–9,985	**0.3665**	**0.1537**	99.81	99.97	97.61	23.3
MIRRA, 1000 steps	**9,988**	0.3824	0.1613	**99.88**	99.97	97.61	–
FragFM, 250 steps	9,878	0.5296	0.3782	98.78	99.97	99.22	81.5
FragFM, 500 steps	9,914	0.5281	0.3824	99.14	99.98	99.23	132.4
FragFM, 1000 steps	9,890	0.5167	0.4010	98.90	99.97	**99.27**	270.6

Every row scores the valid subset of the first 10,000 attempts against all 24,887 references, with failures not replaced. MIRRA at 250 steps is a per-metric median of three sampling draws (valid counts 9,981/9,985/9,980); all other rows are single draws. Times are medians of four 4,096-attempt repetitions, except the FragFM 250- and 1000-step extensions (three repetitions); they cover generation, transfer, assembly and strict acceptance, under the compute setup of Appendix B.2. Quality and timing use separate populations. FragFM500 timing is from the 250-versus-500 study; the separately matched 25-versus-500 study appears in Table 5. At the shared 250- and 1000-step settings MIRRA has lower FCD and NSPDK; no matched 1000-step MIRRA timing is reported. FragFM has higher novelty at all reported settings.

Table 5: **Complete matched generation-timing records.** Every repetition behind the timing statements in the main text, with the median used in each case.

Arm	Attempts per rep	Rep 1 (ms)	Rep 2 (ms)	Rep 3 (ms)	Rep 4 (ms)	Median (ms)
Matched 25 vs. 500 steps						
MIRRA, 25 steps	4,096	26.31	15.77	15.61	15.04	**15.69**
FragFM, 500 steps	4,096	267.43	144.19	122.19	124.86	**134.53**
Matched 250 vs. 500 steps						
MIRRA, 250 steps	4,096	23.40	22.90	23.24	24.48	**23.32**
FragFM, 500 steps	4,096	132.20	131.52	132.67	136.39	**132.43**
FragFM 250-step extension						
FragFM, 250 steps	4,096	81.54	79.19	99.77	–	**81.54**
FragFM 1000-step extension						
FragFM, 1000 steps	4,096	323.44	264.70	270.58	–	**270.58**

Milliseconds per attempted molecule, covering generation, transfer, single-policy assembly and strict acceptance after warmup, and excluding loading, warmup, output writing and distribution scoring. Within each study both arms share the same GPU and batch size 256, in balanced order. MIRRA and FragFM retain different pinned software stacks, and timing repetitions are not training seeds or quality replications.

C.3 Metric conventions

FCD is sensitive to population size and molecular representation. In a separate preprocessing sensitivity analysis, a stereo-stripped training population against the reference scores 0.3361, compared with 0.1881 for a stereo-retaining population. Passing molecules through a simplified element-only graph reconstruction raises the training score to 1.1183, close to the published 1.13 anchor; the same transformation gives 1.1118–1.1332 for the historical MIRRA checkpoints. This transformation changes approximately one third of training molecules and does not reproduce every baseline's charge and component-selection procedure. It explains the scale sensitivity, rather than making the different raw FCD values a controlled ranking. NSPDK's element/bond labels give a much more stable training anchor, 0.1096×10^{-3}, near the published 0.10×10^{-3}.

Our NSPDK implementation follows GDSS: atom-symbol and integer bond-order labels without kekulization; EDeN complexity four, discrete features, 16-bit hashing, inner/global normalization and a fixed Python hash seed. The biased linear-kernel MMD^2, including diagonal terms, is the squared distance between mean feature vectors. This complements ChemNet features without being a statistically independent test.

Table 6: **Fidelity and measured cost against the released FragFM checkpoint.** Both measured MIRRA settings reach lower FCD and lower median generation time than the three tested released settings.

Model	Steps	Median (s)	FCD ↓	Projected FCD ↓
MIRRA	25	17.09	0.1792	0.1814
MIRRA	250	27.83	0.1584	0.1585
Released FragFM	100	44.74	0.4220	0.3071
Released FragFM	250	78.66	0.3261	0.2228
Released FragFM	500	136.54	0.3356	0.2255

Seconds per 1,024 raw attempts including assembly, as the median of four repetitions; the two FCD columns retain the molecular scoring conventions before and after the shared charge projection. Each quality point is one separate population, so the timing repetitions are not error bars. The released and common-cohort checkpoints use different training splits.

C.4 Earlier common-data and released-checkpoint comparisons

The historical common-cohort experiment uses 25,600 raw attempts, first 25,000-valid FCD prefixes and the 24,887-entry stereo-stripped reference. MIRRA250 and locally trained FragFM500 score 0.1584 and 0.3744, a 57.7% reduction; NSPDK is 0.1131 and 0.3272 ($\times 10^3$). This is the earlier result mentioned in the main text, not the current 10,000-attempt estimate. Validity favors MIRRA; uniqueness and novelty favor FragFM. All 16 declared assembly, step-count and valid-prefix contrasts favor MIRRA on both distances. A common charge projection gives FCD 0.1585 versus 0.2875, retaining a 44.9% advantage. The 25-step extension gives FCD 0.1792 with prior assembly.

The authors' released FragFM checkpoint is a separate comparator trained on a split containing 93.9% of the evaluation-reference rows. Table 6 reports its tested operating points and the common-cohort MIRRA checkpoint. Each time is the median of four 1,024-attempt repetitions at batch 256 on one allocation, in rotated order after 256 warmup attempts. Both methods include initialization, sampling, transfer and decoding; loading, writing and scoring are excluded. Their decoders and pinned software stacks remain different.

C.5 Consecutive molecular samples

Figure 7 shows the first 12 outputs of one 250-step benchmark run in generation order, alongside 12 uniformly drawn training molecules. All 12 are chemically valid and connected. No filtering, deduplication or manual selection was applied. Descriptor means use the first 48 outputs of this run and 48 uniformly drawn training molecules; these examples supplement the quantitative tests.

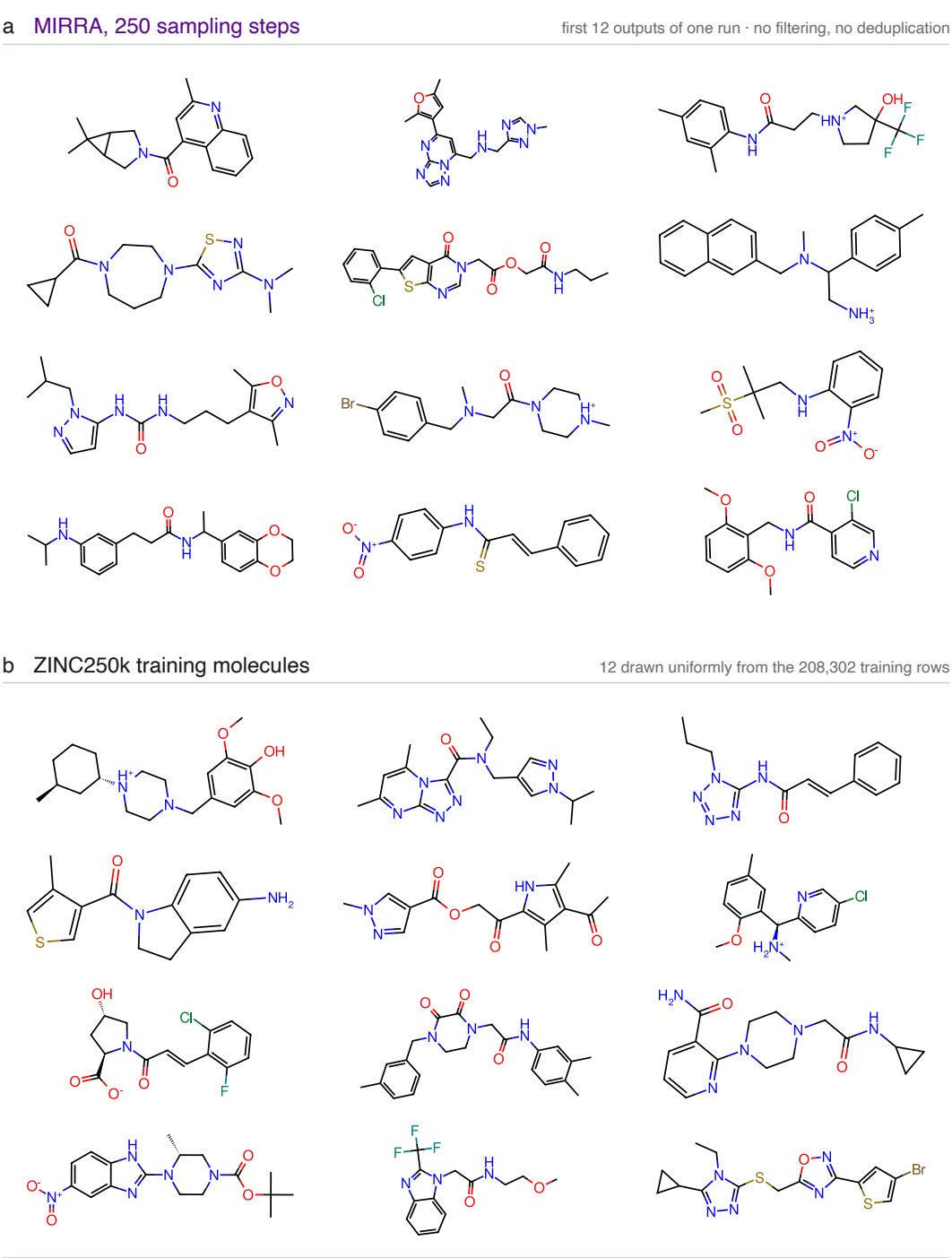

Figure 7: **Consecutive MIRRA outputs beside training molecules.** (a) The first 12 outputs of one 250-step run of the benchmark checkpoint, shown in generation order without filtering, deduplication or manual selection. All 12 are chemically valid and connected. (b) 12 molecules drawn uniformly at random, without replacement, from the 208,302 training rows. The strip beneath reports descriptor means over the full populations these are taken from: the first 48 outputs of the same run and 48 uniformly drawn training molecules. Both panels are small samples shown for qualitative inspection; the distributional evidence is in Table 4.

D FULL-VOCABULARY EXECUTION AND CANDIDATE RESTRICTION

D.1 MATCHED EXPERIMENT BEHIND FIGURE 3

This comparison uses the current 37,821-entry benchmark checkpoint, 250 steps, and the 10,000-attempt quality protocol. The three arms share weights, source laws and assembly: dense full-vocabulary evaluation; scheduled full-vocabulary evaluation; and scheduled candidate-only evaluation. The last draws 1,536 distinct identities without replacement from the training prior per scheduled round and adds current graph identities, so actual set sizes can exceed 1,536. It gathers and projects only candidate columns at nonterminal updates and renormalizes over that set. It uses neither retraining nor a proposal correction. Its sampled terminal state follows the restricted transition, while assembly receives complete terminal posteriors in every arm.

Three quality draws per arm give the fidelity ranges in Table 7. Three separate 1,024-attempt timing runs rotate arm order on the same GPU at batch 128 after warmup, including candidate construction, transfer and assembly. Scheduling reduces median time from 131.56 to 43.74 ms per attempt at essentially unchanged FCD. Candidate scoring uses less peak memory (0.556 versus 0.853 GiB), but takes 46.18 ms and raises FCD to 0.9469. Its higher novelty accompanies a shifted distribution.

D.2 COMPOSITION SHIFTS OCCUR BEFORE ASSEMBLY

A separate replication on three historical models with 38,951 entries uses one 25,000-valid population per arm and model, with full terminal assembly scores. Restriction raises mean FCD from 0.1500 to 0.6925, lowers mean heavy-atom count from 23.07 to 21.77 and molecular weight from 331.1 to 311.7, while raw validity remains above 99.7%. Ten of eleven descriptor distances worsen; logP is the exception. These populations do not replace the current benchmark scores above.

A paired decoder replay uses 512 attempts per arm and model. On identical both-valid pairs, restricted-minus-full heavy-atom differences average -1.145 at the terminal fragment state, -0.016 from slot projection and $+0.058$ from assembly, summing to -1.103 in the final molecule. The raw terminal difference is -1.173. The shift therefore precedes assembly; this stage accounting is not an additive decomposition of the FCD gap.

At a fixed incoming state and realized candidate set S, write $m = p(S)$ and $p^S = p(\cdot \mid S)$. Then

$$D_{\mathrm{TV}}(p, p^S) = 1 - m, \qquad D_{\mathrm{TV}}(K, K^S) = \alpha_k(1 - m). \tag{11}$$

For a fragment property h, with posterior means μ_S and μ_{S^c},

$$\mathbb{E}_{p^S} h - \mathbb{E}_p h = (1 - m)(\mu_S - \mu_{S^c}). \tag{12}$$

Across 4,114 instrumented refreshed-node rows per model, mean omitted posterior mass is 4.75–5.16%, and expected heavy-atom count shifts by -0.195 to -0.215 per refresh. Rows share graphs, candidate sets and trajectories, so they are not independent replicates; state feedback prevents summing these local shifts into terminal FCD.

D.3 AN ORACLE EXAMPLE: SMALLER STEPS DO NOT REMOVE PROPOSAL BIAS

Consider one coordinate with $V \geq 2$ identities, strictly positive source and target both p, and the linear conditional path $q_t(x \mid y) = (1 - t)p(x) + t\mathbf{1}[x = y]$. Its marginal is p for every t, and its exact Bayes endpoint posterior is $p_{1|t}(v \mid x) = (1 - t)p(v) + t\mathbf{1}[v = x]$. At each refresh, independently draw one candidate from a fixed strictly positive distribution ν, add the current identity, and renormalize this posterior over the resulting set.

Proposition. All candidate-averaged transition kernels share the invariant distribution $\pi_i = p_i\nu_i / \sum_j p_j\nu_j$. Their time-one endpoint equals p on every uniform grid, and in the continuous-time limit, if and only if ν is uniform. Full-support refreshes preserve p on every grid.

Proof. A transition from i to $j \neq i$ requires drawing candidate j. The retained posterior mass is $d_{ij}(t) = t + (1 - t)(p_i + p_j)$, giving the averaged kernel $K_k = I + G_{t_k}/T$, where

$$G_t(i, j) = \frac{\nu_j p_j}{d_{ij}(t)} \quad (i \neq j), \qquad G_t(i, i) = -\sum_{j \neq i} G_t(i, j). \tag{13}$$

Table 7: **Exact refresh scheduling preserves full-support quality; candidate restriction changes it.** One benchmark checkpoint at 250 steps.

Setting	FCD ↓	NSPDK ↓	Valid (%)	Unique (%)	Novel (%)
Dense, full vocabulary	0.3680	0.1543	99.81	99.97	97.62
Refresh, full vocabulary	0.3665	0.1537	99.81	99.97	97.61
Refresh, sampled subset	0.9469	0.8780	99.76	99.67	98.14

End-to-end timing and peak allocated memory

Setting	ms/attempt, median [range]	Peak allocation (GiB)
Dense, full vocabulary	131.56 [125.52, 191.77]	1.364
Refresh, full vocabulary	43.74 [42.64, 52.43]	0.853
Refresh, sampled subset	46.18 [44.23, 48.70]	0.556

Quality entries are per-metric medians of three 10,000-attempt draws against 24,887 references; NSPDK is $\text{MMD}^2 \times 10^3$. FCD ranges are 0.3510–0.3741, 0.3514–0.3745 and 0.9057–0.9690 in row order. Timing uses three separate 1,024-attempt repetitions at batch 128, including assembly and acceptance; peak allocation is the maximum across repetitions. These timings have a different batch and execution scope from the between-generator study.

Symmetry of d_{ij} gives detailed balance: $\pi_i G_t(i,j) = \pi_j G_t(j,i)$ for every t. Thus $\pi K_k = \pi$. All entries of K_k are positive, including the diagonal, since drawing the current identity forces a stay. The strict log-sum inequality gives $D_{\mathrm{KL}}(\mu K_k \| \pi) < D_{\mathrm{KL}}(\mu \| \pi)$ whenever $\mu \neq \pi$. If ν is nonuniform, $p \neq \pi$, so a trajectory starting at p cannot return to p after any positive number of steps. If ν is uniform, $\pi = p$ and every kernel preserves it.

The rates in Equation 13 are bounded and smooth on $[0,1]$, so grid refinement converges to $\dot{\mu} = \mu G_t$. With $r_i = \mu_i/\pi_i$, its relative entropy satisfies

$$\frac{d}{dt} D_{\mathrm{KL}}(\mu \| \pi) = -\frac{1}{2} \sum_{i \neq j} \pi_i G_t(i,j)(r_i - r_j)(\log r_i - \log r_j), \tag{14}$$

which is strictly negative away from π. Hence the time-one limit also differs from p for nonuniform ν. Finally, $p\, p_{1|t} = p$ implies preservation under full-support refreshes. This proves the proposition.

For a prior proposal $\nu = p$, the common invariant is proportional to p_i^2; it is generally *not* the finite-time endpoint. This result characterizes the singleton oracle family, not arbitrary candidate-set sizes or molecular graphs. Uniform proposals need not remove bias in those broader settings.

Two-identity closed form. Set $p = \nu = (a, 1-a)$. If the candidate equals the current identity no change is possible; otherwise both identities are available. Since $\alpha_k(1 - t_k) = 1/T$, averaging over the candidate gives the same kernel at every step:

$$K^S = \begin{pmatrix} 1 - (1-a)^2/T & (1-a)^2/T \\ a^2/T & 1 - a^2/T \end{pmatrix}. \tag{15}$$

Writing $c = a^2 + (1-a)^2$ and $\pi = a^2/c$, the probability of the first identity after all T steps is

$$\Pr(X_T = 0) = \pi + (a - \pi)(1 - c/T)^T \;\longrightarrow\; \pi + (a - \pi)e^{-c}. \tag{16}$$

For $a = 0.8$, the restricted endpoint tends to $(0.8696, 0.1304)$, a total variation of 0.06965 from the target $(0.8, 0.2)$. Uniform candidate proposals preserve the target, and reversing the proposal weights reverses the direction of the mass shift. Restriction therefore need not bias every proposal, but time refinement alone does not generally remove its bias.

E Chemical alignment and identity corrections

E.1 Retrained output-head controls

Table 2 fixes input features, backbone, full categorical support, training budget and assembly, while changing the output predictor. All models use the 38,951-entry vocabulary and matched recipe in

Table 8: **Chemical alignment improves fidelity in three independently trained pairs.** Primary 250-step, prior-assembly comparison.

Training pair	Aligned FCD	Shuffled FCD	Reduction (%)
A	0.1561	0.1757	11.19
B	0.1543	0.1727	10.63
C	0.1551	0.1841	15.74

Each pair uses one generation draw per arm. Mean NSPDK$\times 10^3$ is 0.1117 ± 0.0073 aligned versus 0.1368 ± 0.0156 shuffled at 250 steps, and 0.1121 ± 0.0079 versus 0.1278 ± 0.0110 at 1000 (sample SD over models). Both distances favor alignment in all three primary pairs. Execution is matched within each pair; pooled spread also includes hardware and execution differences across pairs.

Appendix B.2; all displayed head distances use the first 25,000 valid outputs and the stereo-stripped reference. These experiments test learning, separately from the fixed-weight interventions below.

The broad shuffle replaces output row s_v by $s_{\pi(v)}$ within positive- and zero-prior strata. It preserves the feature multiset, residual correction, fixed priors and all 27.5M parameters, while breaking feature-to-label and feature-to-frequency associations. Correct alignment lowers primary FCD by 10.63–15.74% across three independently trained pairs (Table 8). The narrower shuffle preserves exact attachment signatures and log2 frequency bins, changing fingerprints for 37.32% of occurrence mass; exact within-bin frequencies are not matched. Its FCD 0.1655 versus 0.1561 aligned gives a 5.72% reduction at 250 steps, with an 8.25% reduction at 1000. Both distances favor alignment in all eight assembly, step-count and valid-prefix contrasts.

Conventional head and checkpoint selection. A $384 \to 264 \to 38{,}951$ ReLU MLP with biases and the same fixed log-frequency offset matches the parameter count within 0.02%. Structural inputs, backbone initialization, training streams and budget are held fixed; output architecture, initialization and score rank change. Minimum EMA validation loss selects update 40,000 for this head and 249,900 for MIRRA. The selected FCDs are 0.4001 versus 0.1561; at the common final update they are 0.1793 versus 0.1561. At 1000 steps, final FCD is 0.1695 versus 0.1500. The main text uses the common-final comparison. An additional weight-decay-0.1 control improves the conventional selected FCD to 0.2625, but worsens its final FCD to 0.2077. MIRRA stays ahead under both policies. The conventional head retains slightly higher novelty (98.05% versus 97.81% at final EMA, 250 steps). This is one trained comparator, not an optimized architecture search or exact retraining of MIRRA with its correction removed.

Reference overlap and representation checks. Removing 285 exact optimizer-training identities from the matched-control reference leaves 24,602 molecules and preserves every comparison's direction: primary FCD is 0.1587 aligned versus 0.1776 shuffled. This checks exact overlap, not near-duplicate memorization. Separately, four common-cohort fragmentations cross implicit/explicit-hydrogen cut selection with all/single-bond BRICS cuts. Their 250-step prior-assembly FCDs span 0.1537–0.1640. The explicit-H, single-bond variant reproduces FragFM's mean training fragment count (6.124) and scores 0.1640 versus 0.3744 for historical FragFM500. Vocabulary, capacity and graph size change together; this checks robustness to representation choices rather than isolating fragment count.

E.2 Where identity corrections affect generation

The current benchmark intervention independently enables the learned identity correction during the generation trajectory and at final assembly. Within a trajectory, the two assembly conditions share terminal states, edge posteriors and paired assembly randomness; only terminal node scores change. Three 10,000-attempt draws per condition use the current reference and are summarized in Table 9.

Holding assembly correction on, enabling trajectory correction reduces FCD by 95.1–95.4% across the three primary draws. Assembly-only correction raises validity by about eight percentage points without recovering FCD. With full trajectory correction, assembly effects on FCD are small and mixed. Repeating the experiment on a second GPU type with fresh sampling draws retains the dominant trajectory effect. These comparisons measure reliance of fixed coadapted weights; the independently trained heads above address the separate learning question.

Table 9: **Corrections during generation dominate the crossed intervention on both devices.** The assembly effect is smaller and metric-dependent.

Correction active at	FCD, median [range] ↓	NSPDK ↓ ($\times 10^3$)	Valid (%)	Unique (%)	Novel (%)
RTX6000					
Neither	7.4210 [7.4134, 7.4297]	16.8574	80.03	96.87	98.59
Assembly only	7.6384 [7.5242, 7.6598]	13.9544	88.10	89.52	98.54
Generation only	0.3675 [0.3538, 0.3726]	0.1600	99.68	99.97	97.65
Generation + assembly	0.3665 [0.3514, 0.3745]	0.1537	99.81	99.97	97.61
H100					
Neither	7.4976 [7.4637, 7.6855]	17.2978	80.51	96.56	98.64
Assembly only	7.4563 [7.4522, 7.7772]	13.5305	88.03	89.59	98.55
Generation only	0.3696 [0.3659, 0.3726]	0.1592	99.70	99.95	97.57
Generation + assembly	0.3734 [0.3713, 0.3735]	0.1614	99.81	99.95	97.48

Each device uses three 10,000-attempt draws per condition from the same checkpoint at 250 steps. All other quality entries are per-metric medians; FCD uses the common 24,887-entry reference. The H100 rows repeat inference, not training. Assembly-only versus neither, and both versus generation-only, need not have the same sign for every metric.

A separate replication uses three historical checkpoints, 4,096 attempts per cell and the first 3,000 valid molecules against an 11,228-molecule internal reference. With full assembly scoring fixed, enabling trajectory corrections decreases FCD from 7.8814 to 0.6264, from 8.2429 to 0.6163 and from 8.3954 to 0.6265 across the three models, a 92.1–92.5% reduction. At first-1,000-valid prefixes and under the alternate assembly rule, every trajectory contrast also favors correction. These values use a different reference and sample size from the current benchmark; they provide a training-checkpoint replication of the intervention, not another benchmark score.

E.3 What the correction adds to the output probabilities

At a fixed contextual state, write the base logit as $b(v) = as_v^\top g(z) + b_v$ and its learned correction as $r(v) = u_v^\top Dz$. Let $B = \mathrm{softmax}(b)$ and $F = \mathrm{softmax}(b + r)$. The attachment groups partition the vocabulary by the multiset of typed attachment sites; the 38,951-entry ZINC vocabulary has 452 such groups. These groups need not have identical structural feature vectors. Define

$$A_g = \log \sum_{v:G(v)=g} e^{b(v)}, \qquad C_g = \log \sum_{v:G(v)=g} e^{b(v)+r(v)}, \qquad \Delta_g = C_g - A_g. \tag{17}$$

Then $\Delta_g = \log \mathbb{E}_{v\sim B(\cdot|g)} e^{r(v)}$ and

$$F(g) = \frac{B(g)e^{\Delta_g}}{\sum_h B(h)e^{\Delta_h}}, \qquad F(v \mid g) = B(v \mid g)e^{r(v)-\Delta_g}. \tag{18}$$

These identities follow by summing the unnormalized logits within each group. A single identity correction therefore induces both a group reweighting and a conditional identity reweighting. The group-only intervention uses logits $b(v)+\Delta_{G(v)}$; identity-only uses $b(v)+r(v)-\Delta_{G(v)}$. Their normalized probabilities are exactly $F(G(v))B(v \mid G(v))$ and $B(G(v))F(v \mid G(v))$, respectively, in real arithmetic. A correction constant within each group changes only group allocation. A correction with every $\Delta_g = 0$ changes only conditional identity probabilities. In general, both change.

Group log-sums are computed in float64 and returned in model precision; the four probability factors match Equation 18 to numerical precision. The algebra explains the intervention, not an independently trained architecture, and its empirical significance comes from the generation differences in Table 10.

The group-only and identity-only posteriors replace the predictor throughout inference, including generation and terminal assembly. They do not modify only the assembly stage.

Table 10 compares all four arms using the first 3,000 valid outputs from each condition's 4,096 attempts, the common evaluation size supported by all four conditions, without replenishment. Group correction restores most validity, while both factors are needed for the best FCD in this model.

Table 10: **Both within-group and between-group corrections contribute.** A whole-inference intervention in one historical checkpoint.

Correction	Valid / attempted	Validity (%)	FCD ↓
Neither	3,313 / 4,096	80.88	8.2378
Group only	4,047 / 4,096	98.80	4.0742
Identity only	3,266 / 4,096	79.74	4.1570
Both	4,090 / 4,096	99.85	0.6264

All four arms score the first 3,000 valid outputs from 4,096 attempts against the same 11,228-molecule internal reference. This common evaluation size is supported by all four conditions, without replenishment. FCD here is not comparable with the current benchmark or the retrained-head experiment. Both corrections act during generation and terminal assembly.

A separate score-space probe uses the projector from Appendix B.3. In the final aligned checkpoint at $t = 0.5$, on 512 validation molecules excluding exact training identities, removing the outside-span correction loses 1.785 target bits/site (paired molecule-bootstrap 95% interval [1.680, 1.893]) and 27.8 top-1 accuracy points across 2,730 positive-count sites. The loss is 1.794 bits on the 2,669 sites with unique feature rows, showing that the effect is not limited to feature collisions. Removing the inside-span component loses 1.360 [1.141, 1.597] bits/site. Both components contribute in the trained model. The full head correctly classifies none of the 52 zero-positive-count targets; this probe establishes no zero-shot accuracy or independently retrained benefit.

The alignment prediction probe finds an inconclusive positive-count gain of 0.0107 bits/site [-0.0027, 0.0244] after excluding 36 exact training identities. At $t = 0.5$, the final-EMA probe covers 3,036 molecules excluded from checkpoint selection and 16,331 sites before this exclusion, including 16,022 positive-count and 309 zero-positive-count targets. In this full probe population, the group-level gain is 0.0119 [0.0037, 0.0198]. Zero-positive-count targets gain 12.15 bits but reach only 2.91% top-1 accuracy versus 0% shuffled; the rare-minus-common contrast does not establish a rare-identity advantage. Thus the generation benefit does not imply uniformly better per-site prediction. Intervals use 2,000 paired molecule-bootstrap resamples, conditional on one trained pair, without multiplicity adjustment or training-seed uncertainty.

F Transfer to MOSES

F.1 Matched training and evaluation

The fixed design is trained on 1,505,430 optimizer rows and 79,233 internal validation rows, using a 51,879-entry vocabulary and 8,192 validation rows for checkpoint selection. Three matched training runs each train aligned, shuffled and conventional output predictors for 42 epochs (493,962 updates), sharing the split, backbone initialization, source priors, corruption, EMA selection and assembly recipe. Aligned and shuffled heads have 30,816,754 parameters. Shuffling permutes whole output feature rows within positive/zero-count strata. The conventional head is a $384 \to 262 \to 51{,}879$ ReLU MLP with biases and the same fixed log-frequency offset, with 30,831,737 parameters (0.05% difference).

Each checkpoint produces 30,720 raw attempts at 250 and 1000 steps, with paired canonical/prior assembly. The primary population is the first 25,000 raw attempts, prior assembly, scored against all 176,225 official scaffold-test (TestSF) molecules (Polykovskiy et al., 2020). Failures are not replenished. The official 176,074-molecule Test reference, 30,000-raw and 25,000-valid prefixes, canonical assembly and 1000 steps are separate sensitivities. Novelty uses unique valid canonical outputs against official training strings. Validation-selected and final EMA checkpoints are reported separately.

F.2 Results across training runs and checkpoint policies

Table 11 gives every run's primary FCD under both checkpoint policies and both step counts. At 250 steps, selected aligned FCD averages 0.5349 versus 0.5507 shuffled and 0.5589 conventional, reduc-

Table 11: **MOSES: three training runs and both checkpoint policies.** FCD against TestSF; smaller is better.

Head	Run	250 steps		1000 steps	
		Selected	Final	Selected	Final
Aligned	A	0.5282	0.5443	0.5169	0.5245
Aligned	B	0.5622	0.5622	0.5299	0.5299
Aligned	C	0.5142	0.5115	0.5333	0.5251
Mean ± SD		0.5349 ± 0.0247	0.5393 ± 0.0257	0.5267 ± 0.0087	0.5265 ± 0.0030
Shuffled	A	0.5626	0.5504	0.5401	0.5443
Shuffled	B	0.5590	0.5536	0.5435	0.5539
Shuffled	C	0.5306	0.5408	0.5355	0.5237
Mean ± SD		0.5507 ± 0.0175	0.5483 ± 0.0066	0.5397 ± 0.0040	0.5407 ± 0.0155
Conventional	A	0.5700	0.5422	0.5586	0.5437
Conventional	B	0.5475	0.5544	0.5324	0.5325
Conventional	C	0.5591	0.5430	0.5317	0.5445
Mean ± SD		0.5589 ± 0.0112	0.5465 ± 0.0068	0.5409 ± 0.0153	0.5402 ± 0.0067

Each cell uses valid outputs among the first 25,000 raw attempts, without refill, prior assembly and the 176,225-molecule scaffold-held-out TestSF reference. Selected and final EMA policies are not pooled. Means and sample standard deviations are across independent training seeds. All heads use the matched recipe and approximately 30.8M parameters.

tions of 2.9% and 4.3%. The common-final means are 0.5393, 0.5483 and 0.5465. The advantage therefore transfers in mean and is smaller than on ZINC.

At 250 steps, aligned prediction wins two of three selected pairs against each comparator. At final EMA it wins two of three against shuffling and one of three against the conventional head. At 1000 steps it wins three of three final pairs against the conventional head and two of three against shuffling. Both comparators retain higher primary novelty. These reversals distinguish a positive mean effect from a uniform improvement across training runs.

Eight 25,000-molecule training draws score scaffold-test FCD 0.518 ± 0.012 (mean and sample SD), providing the training-data floor for this evaluation.

G Property-conditioned generation

G.1 Adaptation and baseline training

MIRRA adaptation. The conditional split contains 207,114 unique training identities and 10,800 validation identities. The property vector holds training-standardized QED (Bickerton et al., 2012) and logP together with a two-entry indicator naming the requested property; the unrequested property's standardized value and its indicator are both zero. For each of the 12 frozen transformer layers, a separate $4 \to 64 \to 544$ MLP with SiLU activation maps this four-dimensional input to offsets of widths 384/96/64 for the node, edge and graph states, broadcast across valid nodes and edges after that layer's output. The final adapter projection starts at zero and a missing condition multiplies every offset by zero, so the unconditioned computation is preserved exactly. Conditioning changes neither the categorical vocabulary nor the linear probability path.

Training updates only the 428,160 adapter parameters. The pretrained backbone and the output predictor stay frozen and in evaluation mode. Each real and shuffled arm uses matched initial adapters and matched data and corruption streams for 20,000 updates at batch 128 of AdamW with learning rate 2×10^{-4}, weight decay 10^{-12} and EMA decay 0.999, retaining the molecular endpoint cross-entropy objective. Each training example requests one property at random, with 10% condition dropout. Every 2,000 updates, a fixed 2,048-row validation subset selects the adapter by denoising loss. Three real-label adaptations share one frozen parent; two have matched fixed-label-permutation controls. The third adaptation extends the primary result, not the control cohort. Generation multiplies the learned offsets by a selected strength in $\{1, 2\}$; this is adapter scaling, not classifier-free extrapolation between two separately evaluated predictions.

External conditional models. FragFM uses its native noisy property-predictor objective and guidance, with the common-cohort locally trained molecular generator frozen. Separate 6.573M-parameter QED and logP predictors train for 50,000 updates at batch 256 of AdamW with learning rate 5×10^{-4}, a 10,000-update warmup and weight decay 10^{-12}, using the authors' four-layer fragment and graph predictors; fixed validation MSE selects the EMA checkpoints. Graph-only and graph-plus-bag guidance remain separate regimes, both retaining 500 generation steps. Graph DiT trains a separate 157.9M-parameter model per property from scratch: width 1152, six layers, 16 attention heads, 500 diffusion steps, Adam with learning rate 2×10^{-4}, and 100 epochs (161,900 updates) at effective batch 128 and condition-drop probability 0.1, with fixed validation noising on 2,048 rows selecting the checkpoint. Training labels and generation requests use the same training-set standardization. The adaptation-based and from-scratch methods therefore differ in training cost and objective.

Guidance selection. For each property and candidate strength, validation generation drew 512 attempts at each of the three targets, and the setting maximizing their mean raw hit rate was selected, with ties resolved toward the first candidate. Guidance was fixed before final evaluation. MIRRA and Graph DiT tested strengths 1 and 2. FragFM was given its native and variance-scaled pairs: for QED, graph strengths 20 and 1042 with paired bag strengths 1 and 52.09; for logP, graph strengths 20 and 9.814 with paired bag strengths 1 and 0.491. QED selected the higher pair, logP the 20/1 pair, and MIRRA and Graph DiT selected strength 2 for both properties. The selection made on the real MIRRA arm was applied unchanged to its shuffled-label control.

G.2 Targets, metrics and matched timing

Targets are the 10th, 50th and 90th training quantiles (q10, q50 and q90): approximately 0.529, 0.763 and 0.885 for QED, and 0.571, 2.626 and 4.186 for logP. All target tolerances and guidance choices were fixed before final scoring. MIRRA uses three adaptations of one shared pretrained generator rather than three independently pretrained backbones; each external method uses one trained model per property. A hit is a valid output whose property lies within ± 0.05 QED or ± 0.5 logP of the requested value, endpoints included. Raw hit rates divide hits by all 4,096 attempts, so invalid attempts and repeated hits stay in the denominator and in the cost accounting. Distinct novel hits count each hitting molecular identity once and additionally require it to be absent from the conditional training identities; identity ignores stereochemistry, and all methods use the same validity, identity and property checks.

Target-window FCD uses the first 1,000 valid outputs from each eligible cell, including valid misses, rather than a subset selected for hitting the target. It scores them against the first 1,000 held-out reference identities inside the same target window; generated duplicates remain and there is no refill. Conditional identity and reference preprocessing differ from the unconditional protocol, so these values are not numerically comparable with the unconditional ZINC evaluation.

Matched conditional timing protocol. The matched comparison measures every primary 500-step setting: three targets for each property, three MIRRA adaptations, both FragFM guidance regimes and Graph DiT. Checkpoints and validation-selected guidance strengths are those used for the existing quality populations. Each setting has three fresh 256-attempt timing repetitions, with method order rotated across repetitions and each cell run in an isolated process under its model's pinned software at batch 128 after 128 warmup attempts, with four CPU threads. Within a target comparison, all methods run on the same physical GPU.

The primary timer starts before random-state and graph initialization and ends after generation, decoding and the shared property, identity, novelty and hit checks. Guidance, failed assemblies and invalid outputs remain inside this interval. A second endpoint before annotation records generation and decoding alone. Loading, warmup, file writing, FCD scoring, training and guidance search are excluded for every method, and GPU work is synchronized at the generation boundaries. FragFM's timed path executes its strict-only decoder, checked in an untimed warmup to reproduce the strict output of the original paired decoder from identical fine predictions.

For each adaptation or baseline setting s, let U_s be the number of distinct novel target hits in its existing 4,096-attempt quality population, and let $t_{s,r}$ be the annotation-inclusive time for timing

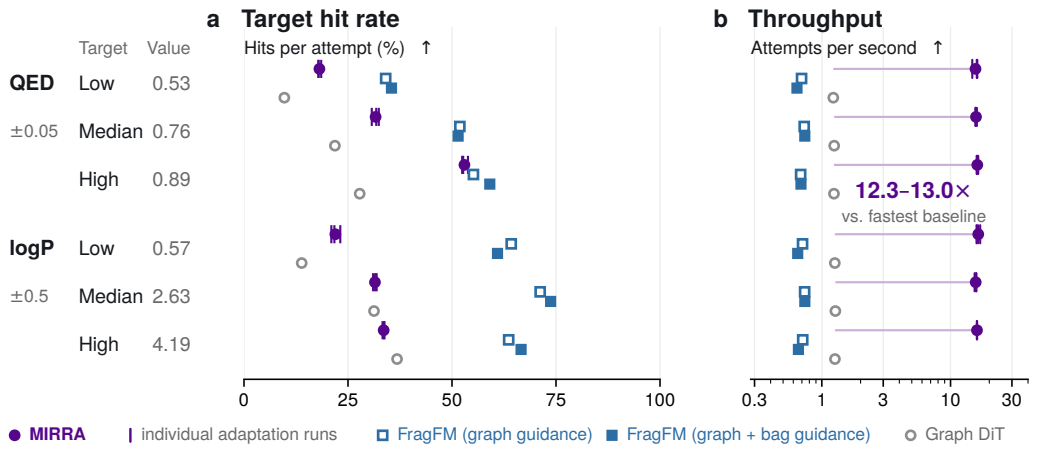

Figure 8: **Hit rate and throughput behind Figure 5a.** (a) Target hits per attempted molecule, as a percentage of all 4,096 attempts. Both FragFM settings reach a higher hit rate than MIRRA at every target. (b) Attempts per second on a logarithmic axis. MIRRA completes 12.3–13.0 times as many attempts per second as the fastest baseline at each target, and 21.3–25.3 times as many as either FragFM setting. The rate in Figure 5a is each run's distinct novel fraction of attempts multiplied by its throughput. MIRRA markers average three property-adaptation runs sharing one frozen backbone; rates and throughputs are averaged separately.

repetition r. The estimated useful rate is

$$q_s = \frac{U_s}{4{,}096}, \qquad \widehat{r}_s = \frac{256 q_s}{\mathrm{median}_{r=1,2,3}\, t_{s,r}}. \tag{19}$$

The reported MIRRA value is the arithmetic mean of its three adaptation-specific rates, not the ratio of averaged yield to averaged time. The mean describes three adaptations of one backbone; it is not the rate of a deduplicated ensemble. Because uniqueness depends on population size, the estimate does not imply a constant long-run discovery rate.

G.3 Complete target grid and cost accounting

Table 12 reports all targets, models and guidance settings across five metrics. Both FragFM settings reach every target more often per attempt than MIRRA does; at the high-QED target FragFM also attains the lower target-window FCD. The claim in Section 5 concerns useful output per unit time; it does not imply higher per-attempt accuracy.

Figure 8 separates hit rate from throughput. MIRRA completes 21.3–25.3 times as many attempts per second as either FragFM setting and 12.3–13.0 times as many as the fastest baseline at each target. That throughput more than offsets a per-attempt hit rate between 0.34 and 0.96 of FragFM's, depending on the target: the gap is widest at low logP and nearly closed at high QED. The resulting rate advantage is 7.8–20.8-fold against the strongest baseline at each target, 7.8–22.5-fold against either FragFM setting and 11.4–24.4-fold against Graph DiT.

Two limits apply to every entry. Yield comes from a fixed 4,096-attempt population while timing comes from separate 256-attempt runs, so these are estimated rates rather than measured outcomes of a fixed-time search; and MIRRA rows average three property-adaptation runs on one shared backbone, which does not sample training-seed variation.

G.4 Label controls and adaptation variability

Table 13 reports the two runs with matched real-label, shuffled-label and unconditional populations. Real conditioning increases mean hit rate at all six targets relative to both controls, with smaller gains at the central targets. Each entry averages two 4,096-attempt populations; these controls are not the three-adaptation averages in the primary grid. Filtering an unconditional population within the same attempt budget cannot increase its number of hits; it only discards misses. Thus the unconditional

Table 12: **Complete property-targeted generation results.** Distinct novel hits per second, hit rate, distinct novel hits per attempt, target-window FCD and validity for every target, model and guidance setting.

Model	Distinct novel hits / second	Hit rate (%)	Distinct novel hits / attempt	Target-window FCD ↓	Valid (%)
QED, low (10th) quantile: target 0.529 ± 0.05					
MIRRA	**2.719**	18.12	0.1739	5.156	99.71
FragFM, graph guidance	0.236	34.06	0.3386	5.330	90.92
FragFM, graph + bag guidance	0.225	35.45	0.3503	5.346	90.21
Graph DiT	0.119	9.69	0.0969	11.614	43.65
QED, median (50th) quantile: target 0.763 ± 0.05					
MIRRA	**4.871**	31.63	0.3089	2.927	99.66
FragFM, graph guidance	0.376	51.88	0.5159	3.434	97.51
FragFM, graph + bag guidance	0.376	51.46	0.5107	3.674	97.34
Graph DiT	0.273	21.85	0.2185	5.307	57.30
QED, high (90th) quantile: target 0.885 ± 0.05					
MIRRA	**8.430**	52.99	0.5224	3.465	99.72
FragFM, graph guidance	0.375	55.15	0.5479	3.284	96.56
FragFM, graph + bag guidance	0.404	59.08	0.5874	3.256	98.44
Graph DiT	0.345	27.81	0.2778	4.678	63.09
logP, low (10th) quantile: target 0.571 ± 0.5					
MIRRA	**3.519**	21.96	0.2145	3.518	99.58
FragFM, graph guidance	0.453	64.23	0.6392	3.636	93.24
FragFM, graph + bag guidance	0.393	60.94	0.6047	4.129	91.11
Graph DiT	0.175	13.87	0.1387	7.231	75.66
logP, median (50th) quantile: target 2.626 ± 0.5					
MIRRA	**4.784**	31.48	0.3050	3.079	99.67
FragFM, graph guidance	0.519	71.19	0.7068	4.873	96.34
FragFM, graph + bag guidance	0.540	73.71	0.7327	4.731	97.02
Graph DiT	0.398	31.25	0.3125	6.412	64.62
logP, high (90th) quantile: target 4.186 ± 0.5					
MIRRA	**5.277**	33.57	0.3288	4.040	99.72
FragFM, graph guidance	0.451	63.62	0.6348	5.385	96.73
FragFM, graph + bag guidance	0.435	66.60	0.6616	4.926	97.39
Graph DiT	0.465	36.79	0.3677	8.605	58.98

All settings use 500 steps, 4,096 attempts and validation-selected guidance. MIRRA rows average three adaptations of one frozen backbone; each baseline regime is one trained model. A hit is valid and inside the target window; distinct novel hits additionally count each non-training identity once. Rates combine each population's useful yield with separate timing repetitions (Equation 19), not a fixed-time search. Target-window FCD scores the first 1,000 valid outputs, including misses, against 1,000 held-out in-window references. It is not comparable with unconditional ZINC FCD.

column also records the raw-budget hit yield of sample-and-filter. It carries no claim about the speed of a separately optimized filtering implementation.

The right-hand columns show each of the three primary adaptations' useful rates, computed separately by Equation 19. These rates show variation across the three adaptation runs for all six targets. They measure adaptation and sampling variability on one frozen parent, not variation across independently pretrained generators.

Table 13: **Condition-label controls and individual adaptation rates.** Control means and primary-run rates are kept separate.

Target	Hit rate (%), two-run mean			Distinct novel hits/s, real labels		
	Real	Shuffled	Uncond.	Run A	Run B	Run C
QED, low	17.91	8.01	8.29	2.540	2.748	2.869
QED, median	31.56	28.50	29.36	4.956	4.828	4.830
QED, high	53.14	26.09	25.67	8.503	8.471	8.317
logP, low	21.36	11.01	10.41	3.494	3.452	3.610
logP, median	31.46	27.98	26.40	4.914	4.765	4.673
logP, high	33.70	15.41	15.41	5.264	5.311	5.255

Left: matched runs A and B, 500 steps and 4,096 attempts per cell. Shuffling permutes condition labels during adapter training; unconditional generation uses the same parent and attempt budget. Right: useful rates for all three real-label adaptations, each computed using its own quality population and three timing repetitions. Run C has no shuffled-label counterpart. Rates are estimates under Equation 19.

H Polymer generation

H.1 Data, representations and baseline adaptations

We use the Open Macromolecular Genome authors' selected all-reactions training set of 99,500 source products across 17 reaction labels rather than the complete OMG database.[1] Canonicalization identifies 99,476 distinct periodic objects. Rotation, reversal and repeated-unit aliases share one split group, so no reserved identity can reappear in training under a relabeling. The split holds 79,580 training, 9,947 validation and 9,949 reserved test objects. Internal generation controls use validation only, including 996 objects with unseen blocks. The external comparison freezes every checkpoint and generated population before scoring the 9,949 reserved test objects (Table 3). This experiment evaluates structural generation on a known linear chain topology. Bulk properties and synthesis feasibility are not evaluated.

It uses two 4,096-attempt draws per model and setting: MIRRA and Graph DiT at 500 diffusion steps, MIRRA additionally at 250 steps from the same checkpoint, and the VAE's native autoregressive decoding, for which no diffusion-step count applies. All 32,768 attempts remain in the denominators; structural distances condition on scorable outputs, whose fractions are reported separately.

Representations and information supplied to each method. MIRRA predicts oriented building blocks on a supplied linear chain, using circular attention and reversal-compatible features; its 14.5M-parameter model trains for 200,000 updates and uses the validation-selected checkpoint. The external models instead generate complete repeat-unit encodings: OMG-VAE emits SELFIES strings (Krenn et al., 2020), and Graph DiT emits atom and bond labels including wildcard attachment atoms. Neither receives MIRRA's fixed chain connectivity. The external comparison therefore measures complete modeling approaches under a common split and scorer, including the advantage of a task-specific representation; it is not an equal-information or equal-training-compute contrast. The paired internal controls reported below use the same representation in both arms.

OMG-VAE adaptation. The 3.952M-parameter model retains the OMG authors' CNN encoder, 152-dimensional latent representation and four-layer bidirectional GRU decoder called one token at a time, with an auxiliary network predicting standardized logP normalized by backbone atom count. Training uses the authors' reconstruction and IMQ-MMD objective with MMD weight 4.345, Adam learning rate 0.002, weight decay 10^{-5}, batch 32 and 60 epochs (149,220 updates); the learning rate falls tenfold at 40 epochs. Autoregressive reconstruction over all 9,947 validation objects under fixed latent draws selects the checkpoint, epoch 45. The SELFIES alphabet and maximum length come from training data only, and the fixed initial attachment token is removed for learning and prepended on decode. Generation draws a Gaussian latent prior with greedy tokens and the au-

[1] `https://github.com/TheJacksonLab/OpenMacromolecularGenome/tree/e4681336c9c7aa1b62ea3f4e31dcf60154d4ba5e`.

thors' nop/second-star stopping rule, so outputs are fresh prior samples rather than reconstructions or optimized latent vectors.

Graph DiT adaptation. The 160.1M-parameter model retains six transformer layers of width 1152 with 16 heads, 500 diffusion steps and Adam learning rate 2×10^{-4}. It trains from scratch for 100 epochs (62,200 optimizer updates) at effective batch 128. The dataset adapter counts wildcard attachment atoms as nodes when estimating graph sizes and no-edge frequencies. A compatibility wrapper restores the original Graph DiT convention of normalized time in training and sampling, leaving transitions, architecture and loss unchanged. Fixed noising on up to 2,048 validation objects selects the checkpoint, epoch 100. The maximum graph size in the training set defines model support. Generation uses the trained size prior and predicts the complete graph without requesting any property.

Periodic adjudication. Ordinary chemical validity requires only a chemically parseable output. Periodic validity is stricter: the repeat unit must be connected, must carry the two required attachment ports, and must canonicalize periodically, with rotation, reversal and repeated-unit aliases collapsing to a single identity. Valid baseline outputs containing building blocks absent from MIRRA's vocabulary are accepted. Structural MMD and descriptor distances use scorable outputs; the raw validity and scorable fractions expose the failures those distances exclude.

H.2 Structural metrics and internal controls

The aligned and permuted models are sampled at 250 diffusion steps (primary) and 1000 steps (sensitivity) with 4,096 raw attempts per population, no rejection or replenishment, matched requested lengths, the selected checkpoints and the full 41,266-token training vocabulary. The reference is the complete 9,947-object validation population, all scorable.

Structural features are ℓ_2-normalized 2,048-dimensional Morgan counts (radius two, chirality enabled), rooted at real atoms of the central unit in a seven-copy chain. For feature vectors x, y, the kernel averages $\exp[-\|x - y\|^2/(2s^2 f)]$ over $f \in \{0.5, 1, 2\}$, with $s^2 = 1.158$ fixed as the median squared distance of 4,096 unequal-index pairs drawn from 2,048 distinct training objects. The unbiased MMD^2 estimator (Gretton et al., 2012) excludes within-population self-pairs and retains duplicates, so it may be negative. The main external table reports MMD^2 in units of 10^{-3}; the smaller internal differences below use 10^{-5}. Descriptor distances are Wasserstein distances over primitive-unit mass, central topological polar surface area (TPSA), and heteroatom, aromatic-heavy, branched-heavy and nonring-single-heavy-bond fractions, each divided by its training-calibration sample standard deviation. These are structural descriptors, not measured bulk polymer properties.

Polymer prediction with shared outputs. Both arms have 14,516,817 parameters and train for 200,000 updates from matching initial weights on the same data. Meaningful output sharing lowers the fixed checkpoint-selection objective at all 100 recorded evaluations. At the final update, meaningful and permuted output features give object-mean endpoint cross-entropies of 2.4632 and 2.5365 nats, a difference of 0.0733 nats or 2.89% of the permuted loss; the selected checkpoints, at 192,000 and 198,000 updates, score 2.4617 and 2.5347. Selection averages each object's token-average loss, so long objects receive no greater weight. These are checkpoint-selection losses on the 1,024 covered validation objects, not likelihoods over the full 9,947-object reference, and the checkpoints are dependent observations from a single training seed.

At 250 steps the permuted-minus-aligned MMD^2 difference is 4.96, and aligned sharing improves five of six descriptor distances; the branched-heavy fraction favors the permuted control (aligned: 0.0359; permuted: 0.0153). Both arms are fully valid, with 98.29% and 97.66% novel draws, whereas a training-resampling control reaches comparably low MMD^2 with zero novelty. At 1000 steps the ordering reverses: aligned MMD^2 is 4.88 against permuted 3.60, aligned sharing is closer only on TPSA and heteroatom fraction, and novelty is 98.36% and 97.56%. The favorable primary point estimate is therefore not robust across these sampling populations, and no ratio or training-seed interval is inferred from such near-zero values.

Fresh sampling repeats at 250 steps. Four further paired draws repeat the 250-step comparison with 4,096 raw attempts per arm, fixed requested lengths and the same 9,947-object reference

Table 14: **Aligned output sharing lowers polymer structural MMD^2 in three of four fresh 250-step draws.** Paired draws of 4,096 attempts per arm against the full validation reference.

Draw	Aligned	Permuted	Difference	Novelty A (%)	Novelty P (%)
1	2.133	2.750	0.6169	98.63	97.27
2	5.900	6.749	0.8484	98.54	97.61
3	8.946	7.810	−1.137	98.76	97.97
4	5.193	8.755	3.562	98.27	97.41
Mean	5.543	6.516	0.9727	98.55	97.56

Unbiased MMD^2 in units of 10^{-5}; positive differences (permuted minus aligned) favor aligned sharing. Novelty A and P are the aligned and permuted novel fractions. Validity is 100% and uniqueness at least 99.85% in every population. The estimator is unbiased and may be negative, so no ratio is formed from these values.

(Table 14). Three of four paired differences favor aligned sharing, and the mean permuted-minus-aligned difference is 0.9727 ± 1.941 (sample standard deviation across draws). Five of six mean descriptor distances favor aligned sharing, with the branched-heavy fraction again favoring the control (aligned: 0.0266; permuted: 0.0227). Mean novelty is 98.55% against 97.56%, favoring aligned sharing in every paired draw, and both arms retain 100% validity. The single reversed draw and the spread of paired differences qualify the positive mean: these are sampling repeats of fixed checkpoints, not training replicates.

These results concern structural generation with a known chain topology. They do not establish uniform gains across sampling settings or independent training seeds, zero-shot whole-model transfer, or coverage of unseen polymer junctions.

H.3 Conventional output head on polymers

A matched comparison replaces only the output head in the circular backbone. The conventional head is a $256 \to 263 \to 41{,}266$ ReLU MLP, plus the same fixed log-frequency prior; its hidden width was chosen by nearest parameter count before training, matching the shared-output model to within 0.03% (14,520,039 versus 14,516,817 parameters). Both arms retain structural inputs, full categorical support, the same 200,000-update recipe, shared initial backbone weights and identical training streams. The validation-selected EMA is primary and the final-update EMA is a separate checkpoint sensitivity; the two policies are never pooled. Each head and policy contributes two paired 250-step draws of 4,096 attempts against the same 9,947-object reference. These are fresh training runs, separate from the aligned/permuted pair above.

Table 15: **The shared output head achieves lower structural discrepancy than a capacity-matched conventional head under both checkpoint policies.** Paired 250-step draws of 4,096 attempts each; unbiased structural MMD^2 in units of 10^{-5}.

Checkpoint policy	Draw	Aligned head	Conventional head
Validation-selected EMA	1	3.243	11.36
	2	0.5184	9.745
	Mean	1.881	10.55
Final-update EMA	1	3.163	13.84
	2	1.081	11.19
	Mean	2.122	12.51

The two draws measure sampling variability of fixed checkpoints, not training-seed uncertainty. Policies are reported separately and never pooled.

Aligned sharing has the lower MMD^2 in all four paired draws (Table 15). The conventional-minus-aligned difference averages 8.671 ± 0.786 under the validation-selected policy and 10.39 ± 0.400 under the final-update policy (mean $\pm$ draw standard deviation). Under the validation-selected policy aligned sharing is closer on all six descriptors; under the final-update policy it is closer on four, with the heteroatom fraction (aligned: 0.0332; conventional: 0.0220) and the nonring-single-heavy-

bond fraction (aligned: 0.0303; conventional: 0.0271) favoring the conventional head. Validity is 100% in all eight populations and novelty ranges from 98.24% to 99.15%.

I Related work

Molecular abstractions and assembly. The junction-tree VAE generates substructure scaffolds and assembles molecules (Jin et al., 2018). HierVAE extends hierarchical generation to larger structural motifs and explicitly predicts their attachments (Jin et al., 2020). Han et al. (2025) use functional groups as coarse-grained graph nodes in a hierarchical autoencoder, with attention-based property prediction for molecular design. NSGGM combines neural proposals with symbolic constraint solving (Geng et al., 2026). FragFM scores individual identities through shared MPNN fragment embeddings and uses learned connection scores with constrained matching (Lee et al., 2026). Its predictor learns density ratios through contrastive fragment bags for a bag-conditioned masked flow. MIRRA uses fixed chemical output features, a separate identity correction and full-vocabulary cross-entropy. Our experiments test the generation consequences of these concrete differences.

Structured output prediction. Our generator uses discrete flow matching's probability paths and endpoint posteriors (Gat et al., 2024). Label Embedding Network learns label relationships and soft training distributions (Sun et al., 2017); deep residual output layers share learned label mappings in language generation (Pappas & Henderson, 2019); LGSA aligns hierarchical label geometry for text classification (Zhang et al., 2026). Building on these established uses of output structure, we test whether chemically aligned sharing improves molecular generation at fixed capacity, input features and probability path. Correction and support interventions then separate additional score expressivity from the choices available during sampling.

Representations and generation paths. GDSS, DiGress, GruM and SID study graph generation through diffusion and iterative denoising (Jo et al., 2022; Vignac et al., 2023; Jo et al., 2024; Boget, 2025). Modular Flows use coupled continuous node dynamics (Verma et al., 2022), while SyCo embeds molecular graphs in synthetic Euclidean coordinates for generation (Ketata et al., 2025). GraphBFN evolves continuous beliefs over graph categories (Song et al., 2025); VBFN couples node and edge beliefs through structured precisions (Xiong et al., 2026). HDFM coarsens and refines graphs (Boget et al., 2026); HOG-Diff uses topological guides in a coarse-to-fine curriculum (Huang & Birdal, 2026). VQ-SAD and GVT learn graph codes before generation (Noravesh et al., 2026; Zheng et al., 2025a); MELD learns masking schedules (Seo et al., 2026), and EGF flows to learned category embeddings (Ma et al., 2026). Gromov–Monge flow matching instead constructs source–target couplings that account for graph permutations (Piening & Wald, 2026). Our output-alignment test fixes representation and path. These experiments concern chemically structured output prediction on a fixed fragment graph; they do not establish an advantage of hierarchical generation paths.

Discrete sampling and property control. Structured discrete diffusion includes both categorical and absorbing-state corruption processes (Austin et al., 2021). Categorical Flow Maps distill categorical flows for generation in fewer steps (Roos et al., 2026). These changes to the learned process complement the execution question studied here: which predictions does a specified sampling transition actually use? For property control, STGG+ combines conditional spanning-tree generation with self-criticism (Jolicoeur-Martineau et al., 2025), while TreeDiff guides diffusion through inference-time tree search (Zhao et al., 2026).

Polymer representations and generation. BigSMILES represents stochastic macromolecular structures (Lin et al., 2019), while G^{2}RINS provides a generative string-and-graph representation of hierarchical polymers (Zaldivar et al., 2026). The Open Macromolecular Genome connects repeat-unit generation to a database built around polymerization reactions (Kim et al., 2023), and Graph DiT studies multi-property-conditioned polymer generation (Liu et al., 2024). Our polymer experiment tests periodic structural generation on a reserved split with known chain topology (Appendix H).