Supplementary Information: Negative Extrapolation for Targeted Motif Suppression in Molecular Generators

Joshua Hesse[1], Julia Westermayr[2,3,4*], Stephan A. Sieber[1*]

[1]TUM School of Natural Sciences, Department of Bioscience, Center for Functional Protein Assemblies (CPA), Chair of Organic Chemistry II, Technical University of Munich, Ernst-Otto-Fischer Strasse 8, Garching bei München, 85748, Germany.
[2]Leibniz Institute of Surface Engineering (IOM), Permoserstraße 15, Leipzig, 04318, Germany.
[3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany.
[4]Institute of Informatics, Leipzig University, Augustusplatz 10, Leipzig, 04109, Germany.

*Corresponding author(s). E-mail(s): julia.westermayr@leibniz-iom.de; stephan.sieber@tum.de;

1 Extended Methods

1.1 Reproducibility structure

All experiments were run from the project repository. The main executable entry points are Makefile targets, which call versioned Python scripts and JSON configuration files. Per-seed sample tables and summaries are retained as the primary analysis inputs; plotting notebooks read these tables, filter to the specified confirmatory seeds, and write both figure files and the plotting data used in each panel. Exploratory method development used seed 11, with seeds 5 and 7 added only for the Transformer scope ablation. Results from these three development seeds were excluded from confirmatory summaries, hypothesis tests, and confirmatory figures. The final manuscript figures are therefore generated from result tables rather than manually entered values.

1.2 RNN and Transformer base generators

We implemented the recurrent neural network (RNN) and Transformer base generators in PyTorch [1] and trained them on the full GuacaMol distribution-learning training set, comprising 1,273,104 molecular SMILES derived from ChEMBL 24 [2]. The corresponding configurations are provided in `configs/guacamol_rnn_base.json` and `configs/guacamol_transformer_base.json`.

The RNN contained a 256-dimensional token-embedding layer, a three-layer gated recurrent unit network with hidden dimension 768 and dropout 0.15, and a linear vocabulary projection [3]. The Transformer contained 256-dimensional token and positional embeddings, four causally masked Transformer encoder blocks with eight attention heads, a feed-forward dimension of 1024 and dropout 0.1, followed by layer normalization and a linear vocabulary projection [4]. Its maximum input sequence length was 192 tokens.

Both models used the same regex-based SMILES tokenizer. Bracketed atom expressions, two-character element symbols, ring indices, bonds, branches, and stereochemical symbols were retained as individual tokens, and padding, beginning-of-sequence, and end-of-sequence tokens were added. The vocabulary was fitted on the GuacaMol training set. Both generators were trained autoregressively to predict the next token using padding-masked categorical cross-entropy and AdamW [5]. The RNN was trained for eight epochs with batch size 512 and learning rate 0.001. The Transformer was trained for eight epochs with batch size 256 and learning rate 0.0005. The fixed eight-epoch budget was selected during base-model development, before downstream objective evaluation, and was held constant across seeds.

Molecules were generated autoregressively by multinomial sampling from the next-token probability distribution, with a maximum generated length of 150 tokens. Sampling temperatures were 0.9 for the RNN and 0.75 for the Transformer; no top-k or nucleus truncation was applied. Confirmatory models were trained independently with seeds 13, 17, 19, 23, 29, 31, 37, 41, 43, and 47. The Transformer scope ablation used development seeds 5, 7, and 11; the remaining single-seed sensitivity and parameter-selection experiments used development seed 11. None of these development runs entered confirmatory summaries or tests.

1.3 Objective-specific editing workflow

For each seed, the saved base checkpoint generated an initial pool of 30,000 molecules. Objective-specific negative and positive sets were selected from valid scored molecules in this pool. Binary motif-removal objectives used negative-set molecules with the target flag equal to one and positive-set molecules with the same flag equal to zero. Selection was balanced by downsampling and capped at 5000 molecules per set. The configured minimum was 500 molecules per set, except for the Transformer charged-motif and assay-interference objectives, which used minima of 450 and 300, respectively. The RNN objective runs used two negative-set fine-tuning epochs and two positive fine-tuning epochs. Transformer objective runs used one epoch for each update. These short updates were chosen to estimate local task vectors while limiting

drift from the base generator; Supplementary Figure S6 evaluates the sensitivity to longer updates.

To assess sensitivity to fine-tuning duration, we additionally evaluated the reactive-liability objective on development seed 11 at 1, 2, 5, and 10 continuous fine-tuning epochs for the RNN, Transformer, and REINVENT prior. Both negative-set and positive-set checkpoints were captured after each requested epoch. Standard negative extrapolation (NE) was calculated from each negative-set checkpoint at $\lambda = 1.0$. To distinguish update magnitude from direction, an epoch-norm NE series globally rescaled each negative-set direction to the established architecture-specific direction norm: two epochs for the RNN, one epoch for the Transformer, and two epochs for the REINVENT prior. Transformer extrapolation retained the final last-block/output scope; the RNN and REINVENT prior used full-model extrapolation. Positive and negative-set fine-tuning updated the full model in every architecture. Each endpoint used 5000 requested samples generated with a common sampling seed. For the RNN and Transformer, a usable molecule was valid, canonically unique, absent from the GuacaMol training set, and free of the reactive liability. Because a definitive training-set novelty reference was unavailable for the external prior, the corresponding definition for the REINVENT prior required validity, canonical uniqueness, and absence of the reactive liability. Unique usable scaffold yield was the number of distinct non-empty Bemis–Murcko scaffolds among these molecules divided by all 5000 requested outputs. This exploratory analysis was not used as a confirmatory replicate.

The quantitative estimate of drug-likeness (QED) control used the bottom 10% of valid base-generated molecules by QED as the negative set and the top 10% as the positive set. Liability-removal objectives used the structural flags listed in Table S1. A molecule was positive for a family if any SMARTS pattern assigned to that family matched its RDKit molecular graph. Exact SMARTS definitions are given in Table S2, and representative highlighted matches are shown in Supplementary Figure S1.

For every random-NE control, a base-model copy was fine-tuned with the same optimizer, epoch count, batch size, and learning rate on a size-matched random subset from the corresponding base-generation pool. Let $\Delta_{\text{bad}} = \theta_{\text{bad}} - \theta_{\text{base}}$ and $\Delta_{\text{rand}} = \theta_{\text{rand}} - \theta_{\text{base}}$. Over all floating-point tensors in the exact parameter scope used by the corresponding NE experiment, the random direction was scaled by

$$s_{\text{rand}} = \frac{\|\Delta_{\text{bad}}\|_2}{\|\Delta_{\text{rand}}\|_2}, \qquad \theta_{\text{random NE}} = \theta_{\text{base}} - \lambda s_{\text{rand}} \Delta_{\text{rand}}. \tag{S1}$$

The norm was computed globally across the scoped tensors, rather than independently per tensor. Norms and scale factors were saved per seed, objective, architecture, and parameter scope. The control for the REINVENT prior used the same trained-random construction; no synthetic Gaussian parameter vector was used in the confirmatory experiments.

1.4 Chemical rationale and interpretation of the structural objectives

The four structural families were study-defined operational objectives, not one-to-one reproductions of a published filter catalog. They were assembled to represent chemically interpretable regions that occur often enough in generated libraries to support paired model-editing experiments. The family-level flags should therefore be interpreted as controllable structural proxies rather than predictions of toxicity, assay failure, or developability. In particular, a SMARTS match is a reason for chemical review, not evidence that a molecule is intrinsically unsuitable.

The reactive family covers electrophilic or intrinsically reactive groups, including acyl and sulfonyl chlorides, isocyanates, strained three-membered heterocycles, aldehydes, Michael acceptors and related activated alkenes, and alkyl halides. Such groups are commonly considered during screening-library triage and covalent-inhibitor design [6, 7]. Their inclusion here does not imply that covalent warheads are universally undesirable: appropriately tuned electrophiles are established pharmacophores in targeted covalent drugs. The objective instead models a task in which unplanned covalent reactivity is to be avoided.

The metal-binding family contains catechol, hydroxamic acid, amidoxime, thiol, dithiocarbamate, and dicarbonyl patterns. These motifs contain donor-atom arrangements capable of metal coordination. Metal-binding groups can create promiscuity or assay sensitivity in an untargeted campaign, but can also be essential pharmacophores for selective metalloenzyme inhibitors [8].

The charged family detects quaternary ammonium, sulfonate, phosphonate, carboxylate, and co-occurring formally positive and negative atoms. These groups are not general medicinal-chemistry liabilities; carboxylates and zwitterions are common in approved drugs. This family is included as a deliberately broad controllability benchmark for a chemically coherent, readily detected region of graph space that may be undesirable in a particular permeability, distribution, or formulation context. No pH-dependent protonation or pK_a prediction was performed, so the endpoint refers only to formal charges in the generated molecular representation.

The assay-interference family contains rhodanine, quinone, azo, nitro, thiourea, and polyphenol patterns. These chemotypes overlap with motifs discussed in the PAINS and frequent-hitter literature and can be associated with redox activity, aggregation, covalent reactivity, optical interference, or other assay-dependent mechanisms [9, 10]. Structural-alert filters have a limited applicability domain and should not be used as automatic proof of interference [11]; here they provide a reproducible motif-suppression objective rather than a biological annotation.

Table S2: Exact study-defined SMARTS patterns used for the four confirmatory structural objectives. A family flag is positive if any listed pattern matches. Pattern names reproduce the identifiers in `neon_molgen/scoring.py`.

Family	Pattern	SMARTS
Reactive motifs	acid chloride	`C(=O)Cl`

Continued on next page

Table S2 – continued

Family	Pattern	SMARTS
	sulfonyl chloride	`S(=O)(=O)Cl`
	isocyanate	`N=C=O`
	isothiocyanate	`N=C=S`
	aldehyde	`[CX3H1](=O)[#6]`
	epoxide	`[OX2r3]1[#6r3][#6r3]1`
	aziridine	`[NX3r3]1[#6r3][#6r3]1`
	michael acceptor	`[C,c]=[C,c][C,S](=O)`
	alpha beta unsaturated carbonyl	`[C,c]=[C,c]C(=O)[#6,#7,#8]`
	alpha beta unsaturated nitrile	`[C,c]=[C,c]C#N`
	nitroalkene	`[C,c]=[C,c][N+](=O)[O-]`
	vinyl sulfone sulfoxide	`[C,c]=[C,c]S(=O)`
	haloacetamide	`[Cl,Br,I]CC(=O)N`
	maleimide	`O=C1NC(=O)C=C1`
	alkyl halide	`[CX4;!$(C(F)(F)F)][Cl,Br,I]`
Metal-binding motifs	catechol	`c([OH])c([OH])`
	hydroxamic acid	`C(=O)N[OH]`
	amidoxime	`C(=N[OH])N`
	thiol	`[SH]`
	dithiocarbamate	`N(C(=S)S)`
	beta dicarbonyl	`C(=O)C[C,N,O,S]C(=O)`
Charged motifs	quaternary ammonium	`[NX4+]`
	sulfonate	`S(=O)(=O)[O-]`
	phosphonate	`P(=O)([O-])[O-]`
	carboxylate	`C(=O)[O-]`
	zwitterion proxy	`[+].[-]`
Assay-interference motifs	rhodanine	`O=C1NC(=S)SC1`
	quinone	`O=C1C=CC(=O)C=C1`
	azo	`[#6]-N=N-[#6]`
	nitro	`[$([NX3](=O)=O),$([N+](=O)[O-])]`
	thiourea	`NC(=S)N`
	polyphenol	`c([OH])cc([OH])`

1.5 Transformer scope ablation

The Transformer scope ablation used metal-binding-motif removal as the representative objective and development seeds 5, 7, and 11. For every seed, all parameter scopes were constructed from the same base checkpoint and negative-set fine-tuning update. We compared full-model extrapolation with exclusion of embeddings and layer-normalization parameters, the final two Transformer blocks plus output head, the final Transformer block plus output head, and the output head alone. Every scope was evaluated at $\lambda \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$ with 10,000 requested samples per scope, lambda, and seed. The final confirmatory Transformer liability experiments used parameter prefixes `transformer.layers.3.` and `output.`, corresponding to the last Transformer block and output head in the project implementation. This development experiment was used only to select an update scope and was excluded from confirmatory hypothesis tests.

Table S1 Structural objectives used in the confirmatory analyses. The interpretation column states the intended experimental use and avoids treating every match as an intrinsic molecular liability.

Objective	Negative-set flag	Operational interpretation
Reactive motifs	`reactive_hit`	Avoid unplanned electrophilic or intrinsically reactive functionality.
Metal-binding motifs	`chelator_hit`	Avoid selected donor-atom arrangements when nonspecific metal coordination is undesirable.
Charged motifs	`charged_motif_hit`	Benchmark suppression of formally charged graph motifs for context-specific design briefs.
Assay-interference motifs	`assay_interference_hit`	Avoid selected structural alerts associated with assay-dependent interference mechanisms.

1.6 REINVENT prior experiments

These experiments used the public classical recurrent REINVENT prior distributed with REINVENT4 at `external/REINVENT4/priors/reinvent.prior`. The ChEMBL-trained checkpoint has a 256-dimensional embedding and three LSTM layers of hidden size 512. For each motif-removal objective and seed, 50,000 baseline molecules were sampled, negative and positive sets were selected from this pool, and each model was evaluated with 10,000 generated molecules. Transfer learning minimized the mean autoregressive SMILES negative log-likelihood with the native REINVENT4 implementation; no activity or motif label was supplied to the model. It used two epochs, batch size 256, learning rate 10^{-4}, randomized SMILES, and unique-molecule sampling. The short two-epoch duration was chosen to estimate a local task vector while limiting drift from the pretrained prior; Supplementary Figure S6 evaluates longer updates. The NE lambda grid was 0.25, 0.5, 0.75, and 1.0. Random NE was derived from a size-matched random transfer-learning set and globally norm-matched to the negative-set transfer-learning direction before extrapolation.

The fixed-budget Base evaluation used a separate 10,000-attempt draw per confirmatory seed (sampling seed equal to the replicate seed plus 1000), scored for all four objectives. Endpoint and intrinsic-diversity comparisons use this evaluation, not the 50,000-molecule training-selection pool. The saved 50,000-molecule pools serve as distributional-distance references and as the source for molecular-weight calibration.

1.7 SemlaFlow architecture and checkpoint

The three-dimensional experiments used the public GEOM-Drugs checkpoint `external/semla-flow/assets/models/geom-drugs/200epochs.ckpt` from SemlaFlow [12]. The checkpoint hyperparameters specify a 384-dimensional model, 12 equivariant layers, 32 attention heads, 128-dimensional edge and message representations, 64 coordinate sets, 18 atom-vocabulary entries, five edge types, and a maximum of 256 atoms. The model jointly represents atom coordinates, atom types, bond types, and formal charges. The public checkpoint was trained for 200 epochs with

scale optimal transport, self-conditioning, and exponential moving-average parameters. Sampling used the exponential moving-average generator, 100 integration steps, logarithmic ODE sampling, categorical noise level one, and zero coordinate-noise standard deviation.

1.8 SemlaFlow objective construction and fine-tuning

Independent base pools of 50,000 generated molecules were prepared for confirmatory seeds 13, 17, 19, 23, 29, 31, 37, 41, 43, and 47. Molecules were scored for the same reactive, metal-binding, charged-motif, and assay-interference flags used in the string-based generator experiments. The joint endpoint `four_liability_hit` was the logical union of these flags.

The negative-set selector assigned unique molecules to liability families and sampled 3000 training and 300 validation molecules per family. This produced 12,000 training and 1200 validation molecules without allowing multi-hit molecules to be duplicated across assigned family quotas. Matched random and liability-free correction sets contained 12,000 training and 1200 validation molecules. All SemlaFlow fine-tuning updates used ten epochs, learning rate 3×10^{-4}, and the full flow-matching model. Fine-tuning retained the native SemlaFlow objective: masked mean-squared error for predicted coordinates together with the native atom-type, bond-type, and formal-charge categorical losses. All four component weights were one. Molecules were not supplied with positive or negative labels; set membership determined which molecular examples contributed to the unchanged generative loss. Derived checkpoints were sampled once and deleted after evaluation to limit storage use.

Standard NE used

$$\theta_{\mathrm{NE}} = \theta_{\mathrm{base}} - \lambda(\theta_{\mathrm{bad}} - \theta_{\mathrm{base}}), \tag{S2}$$

whereas corrected NE (CNE) used the liability-free task vector as a validity-correction direction:

$$\theta_{\mathrm{CNE}} = \theta_{\mathrm{base}} - \lambda\left[(\theta_{\mathrm{bad}} - \theta_{\mathrm{base}}) - (\theta_{\mathrm{free}} - \theta_{\mathrm{base}})\right]. \tag{S3}$$

Thus, standard NE and CNE use the reported λ without additional norm scaling. The liability-free direction was used to cancel changes in the negative-set update associated with continued training on valid chemistry; it was not treated as a separate enrichment objective. SemlaFlow random NE used the trained-random direction defined above and the standard random-NE scaling factor $s_{\mathrm{rand}} = \|\Delta_{\mathrm{bad}}\|_2 / \|\Delta_{\mathrm{rand}}\|_2$.

As a mechanistic control, we also defined a random-corrected direction and the CNE reference direction

$$\begin{aligned} v_{\mathrm{CNE}} &= \Delta_{\mathrm{bad}} - \Delta_{\mathrm{free}}, \\ v_{\mathrm{rand}} &= \Delta_{\mathrm{bad}} - \Delta_{\mathrm{rand}}, \end{aligned} \tag{S4}$$

and matched the global norm of the random-corrected direction to that of the CNE direction using

$$s_{\mathrm{control}} = \frac{\|v_{\mathrm{CNE}}\|_2}{\|v_{\mathrm{rand}}\|_2} = \frac{\|\Delta_{\mathrm{bad}} - \Delta_{\mathrm{free}}\|_2}{\|\Delta_{\mathrm{bad}} - \Delta_{\mathrm{rand}}\|_2}. \tag{S5}$$

The resulting model was

$$\theta_{\text{random-corrected control}}(\lambda) = \theta_{\mathrm{base}} - \lambda s_{\mathrm{control}}(\Delta_{\mathrm{bad}} - \Delta_{\mathrm{rand}}). \tag{S6}$$

Results for this control are reported in Supplementary Section 2.2. The factors s_{rand} and s_{control} therefore serve different purposes: the former matches the random task-vector norm to the negative-set task-vector norm, whereas the latter matches the random-corrected control to the CNE direction. All norms were computed globally over the floating-point tensors in the full-model update scope. The main comparison used $\lambda = 2.5$ and 5000 evaluation molecules per seed and method.

1.9 SemlaFlow parameter selection

SemlaFlow settings were selected in an exploratory sequence using development seed 11 and a Brenk oxygen–nitrogen single-bond proxy objective (Supplementary Figure S5). We first examined whether the negative-set fine-tuning update learned the proxy objective across learning rates and epoch counts (panel a). We then extrapolated the resulting direction over full-model and restricted parameter scopes (panel b). Restricted scopes preserved graph validity but produced little target removal, whereas the full-model direction produced a measurable liability response and was therefore retained. A subsequent lambda screen quantified the trade-off between stronger proxy removal and declining RDKit-valid graph yield under standard NE (panel c). The same screen indicated that subtracting a trained correction direction could recover much of the lost validity, motivating CNE and its random-corrected control in the confirmatory experiment. Finally, the selected full-model operating point was checked once against each of the four medicinal-chemistry liability families before the joint objective was constructed (panel d). The resulting settings—ten fine-tuning epochs, learning rate 3×10^{-4}, full-model extrapolation, and $\lambda = 2.5$—were then transferred unchanged to the ten confirmatory seeds. Development seed 11 and all proxy-objective screens were excluded from confirmatory summaries and hypothesis tests.

1.10 SemlaFlow validity, three-dimensional quality, and efficiency metrics

RDKit-valid graph yield was calculated as the number of sampled outputs that could be sanitized, canonicalized, reparsed, and processed by the descriptor pipeline divided by the original number sampled. PoseBusters was run with the `mol` configuration on available SDF records. The conditional PoseBusters rate shown in the main figure is the intersection of RDKit-valid and PoseBusters-passing outputs divided by the number of RDKit-valid outputs. Missing or unparsable outputs therefore affect graph yield but do not enter this conditional denominator.

Three-dimensional usable yield used the original fixed sampling budget as the denominator and required the intersection of RDKit validity, canonical-SMILES uniqueness, absence of all four liability families, and passage of every PoseBusters check. No external novelty reference was available for the public SemlaFlow checkpoint, so novelty was not included in this endpoint. Scaffold diversity was the number of unique non-empty Bemis–Murcko scaffolds divided by the number of unique three-dimensional usable molecules. Usable scaffold yield divided the same scaffold count by all sampled molecules. Fixed-target efficiency was the number of sequential samples required to observe 2500 unique three-dimensional usable scaffolds.

1.11 Two-dimensional molecular metrics

SMILES validity required successful RDKit parsing, sanitization, canonicalization, and descriptor calculation. Uniqueness was calculated from canonical SMILES among valid outputs. For the RNN and Transformer, novelty required a canonical SMILES not present in the canonicalized GuacaMol training set. Because the complete training sets associated with the public REINVENT prior and SemlaFlow checkpoints were not available in an equivalent canonicalized form, novelty was omitted from their fixed-budget yield intersections rather than estimated against an incomplete reference.

QED was calculated with RDKit `QED.qed`. Molecular weight used `Descriptors.MolWt`; logP used the Wildman–Crippen implementation `Crippen.MolLogP`; TPSA, hydrogen-bond donor and acceptor counts, ring counts, aromatic ring counts, and fraction of sp^3 carbons used the corresponding `rdMolDescriptors` functions. Formal charge was obtained from the sanitized molecular graph. Bemis–Murcko scaffolds were generated with RDKit. For the reported scaffold fractions, empty scaffolds from acyclic molecules were excluded from the distinct-scaffold numerator, while these molecules remained in the corresponding molecule-count denominator. Morgan fingerprints used radius 2 and 2048 bits. Mean pairwise Tanimoto distance was $1 - \overline{T}$ over sampled fingerprint pairs, where $\overline{T}$ is mean binary-fingerprint Tanimoto similarity.

1.12 Distributional metrics

Fréchet ChemNet distance (FCD) was computed with `fcd-torch` from ChemNet embeddings of unique valid canonical SMILES. For the metal-binding experiments, up to 3000 molecules per generated/reference set were used for the RNN and Transformer, and up to 5000 for the REINVENT prior. SemlaFlow comparisons used up to 5000 molecules per set. Molecules were selected without replacement using deterministic comparison-specific seeds, with fewer molecules used when a set was smaller. The Fréchet distance was calculated from the two embedding means and covariance matrices. Distances were calculated both to the saved raw-base reference and, for liability experiments, to the liability-free subset of that base reference.

In the distribution-shift figure, Base denotes draws from the saved reference pool, not the separate fixed-budget REINVENT evaluation. Base-to-raw-base distances compare identical molecule sets and are zero by construction. Sampling variation is assessed separately by the molecular-weight calibration's baseline-resampling control.

Fréchet descriptor distance (FDD) followed the implementation introduced by Özçelik and Grisoni [13]. Each molecule was represented by the five-dimensional vector (logP, MW, HBD, ring count, TPSA). Features were linearly scaled using fixed minima $(-3, 0, 0, 0, 0)$ and maxima $(10, 1000, 10, 10, 250)$ before applying the same Gaussian Fréchet formula used for FCD.

Molecular-weight calibration used the saved base-generated pools from the metal-binding experiments. For each seed, query molecules were drawn from either the unfiltered pool or one of seven molecular-weight quantile bands. Reference molecules were drawn from the unfiltered pool. Sampling was without replacement within each set, but query and reference draws could overlap; they were not separate generation

runs. In the primary calibration, every query and reference within a generator/seed used the same count, determined by the smallest band and capped at 5000. Counts ranged from 916–930 for the RNN, 903–952 for the Transformer, and 4905–4916 for the REINVENT prior (Supplementary Figure S7).

We also retained two sample-size sensitivity calculations: the original procedure with up to 5000 query molecules and 5000 reference molecules, and a pair-matched procedure retaining each original query and restricting its reference to the same count. These calculations illustrate the sample-size dependence of the distances, especially FCD. The primary calibration equalizes counts across molecular-weight bands, but does not match the sample counts of all model comparisons. The controls therefore provide qualitative context for chemical-space restrictions, not universal or sample-size-independent thresholds for FCD or FDD.

1.13 Statistical summaries

Endpoint comparisons used paired seed-level values for seeds 13, 17, 19, 23, 29, 31, 37, 41, 43, and 47. Each seed defined one experimental replicate; generated molecules were not treated as independent replicates. Pairing was preserved because methods within an architecture and objective used the same starting checkpoint and seed-specific generated pool. For each comparison, we report the mean paired difference and a two-sided 95% Student-t confidence interval for that mean difference. Hypothesis tests used an exact two-sided paired sign-flip test: all 2^{10} assignments of signs to the ten paired differences were enumerated, and the p-value was the fraction whose absolute mean was at least as large as the observed absolute mean.

Holm correction was applied separately within each prespecified generator–endpoint family. For each string-based generator, a family comprised the four motif-removal objectives and the three comparisons of NE with base generation, positive fine-tuning, and norm-matched trained-random NE, giving 12 tests per endpoint. Each QED family contained the three comparator tests for one architecture and endpoint. Each SemlaFlow family contained the five comparisons of CNE with base generation, positive fine-tuning, standard NE, random NE, and the norm-matched random-corrected control for one endpoint. Because ten pairs permit only 2^{10} sign assignments, the exact p-values are discrete; repeated adjusted values therefore reflect the test resolution and common effect direction rather than duplicated observations.

The target-hit endpoint is the fraction of valid molecules matching the objective-specific structural flag. For the RNN and Transformer, usable yield additionally requires canonical uniqueness and novelty relative to the GuacaMol training set. For the public REINVENT prior, the complete canonical training set was unavailable, so the corresponding fixed-budget yield requires validity, canonical uniqueness, and liability absence but not training-set novelty. SemlaFlow three-dimensional usable yield additionally requires passage of all PoseBusters checks. Fractional endpoint differences and confidence limits are reported in percentage points. Scope screens, lambda-response analyses, fine-tuning epoch sensitivity, and distributional-distance analyses are descriptive and were not included in confirmatory hypothesis tests.

Primary motif-removal comparisons, QED-control comparisons, and SemlaFlow comparisons are reported in Supplementary Tables S3, S4, and S5, respectively. Paired

effect estimates for the three string-based generators are visualized in Supplementary Figure S10; the corresponding SemlaFlow comparisons are shown in Supplementary Figure S12.

1.14 Computational resources

Computations were performed on a system with an AMD Ryzen Threadripper PRO 5995WX CPU (64 cores and 128 threads), 1024 GB RAM, and an NVIDIA RTX 4090 GPU with 24 GB VRAM.

2 Supplementary Results

2.1 Confirmatory paired comparisons

Complete paired comparisons for the QED control, the string-generator motif-removal experiments, and the SemlaFlow experiment are reported in Supplementary Tables S4, S3, and S5, respectively. The corresponding effect estimates are visualized in Supplementary Figures S10 and S12.

2.2 Random-correction control

To test whether CNE required a correction direction learned specifically from motif-free molecules, we compared it with a control constructed from a size-matched random set of valid generated molecules. The random-corrected direction was globally norm-matched to the CNE direction. It restored RDKit-valid graph yield to 87.4% and reduced the joint motif-hit frequency to 3.4%, compared with 88.3% and 2.2%, respectively, for CNE. Relative to the random-correction control, CNE further reduced motif frequency by 1.21 percentage points and increased three-dimensional usable yield by 1.93 percentage points; both differences remained significant after Holm correction ($p_{\mathrm{Holm}} = 0.0098$; Supplementary Figures S11–S12 and Supplementary Table S5). These results are consistent with a generic trained correction direction accounting for most of the validity recovery, while selection of motif-free molecules provided a smaller additional benefit.

Table S3 Primary paired comparison of NE with positive fine-tuning. Effects are NE minus positive fine-tuning in percentage points with 95% confidence intervals. Negative target-hit effects and positive usable-yield effects favor NE. Exact two-sided sign-flip p-values are Holm-adjusted within each generator and endpoint across all twelve prespecified comparisons.

Generator	Objective	Target-hit difference; p_{Holm}	Usable-yield difference; p_{Holm}
RNN	Reactive	-3.94 [-4.33, -3.55]; 0.0234	-3.42 [-4.78, -2.06]; 0.0234
RNN	Metal-binding motif	-7.80 [-8.53, -7.08]; 0.0234	-4.25 [-5.36, -3.15]; 0.0234
RNN	Charged motif	-0.32 [-0.47, -0.18]; 0.0234	-6.56 [-8.55, -4.58]; 0.0234
RNN	Assay interference	-0.97 [-1.25, -0.69]; 0.0234	-6.23 [-7.56, -4.90]; 0.0234
Transformer	Reactive	-2.17 [-2.43, -1.91]; 0.0234	+13.23 [+9.52, +16.95]; 0.0234
Transformer	Metal-binding motif	-5.41 [-5.99, -4.83]; 0.0234	+6.89 [+5.42, +8.37]; 0.0234
Transformer	Charged motif	-0.91 [-1.13, -0.69]; 0.0234	+35.65 [+25.28, +46.02]; 0.0234
Transformer	Assay interference	-1.03 [-1.21, -0.85]; 0.0234	+22.60 [+15.80, +29.40]; 0.0234
REINVENT prior	Reactive	-4.43 [-4.65, -4.21]; 0.0234	+4.68 [+4.49, +4.86]; 0.0234
REINVENT prior	Metal-binding motif	-4.99 [-5.29, -4.69]; 0.0234	+4.88 [+4.55, +5.21]; 0.0234
REINVENT prior	Charged motif	-2.59 [-2.71, -2.47]; 0.0234	+2.92 [+2.79, +3.06]; 0.0234
REINVENT prior	Assay interference	-3.85 [-4.06, -3.64]; 0.0234	+3.93 [+3.72, +4.14]; 0.0234

Table S4 Paired QED-control comparisons. Effects are NE minus comparator with 95% confidence intervals. QED ≥ 0.9 effects are percentage points; mean-QED effects are in QED units. Exact two-sided sign-flip p-values are Holm-adjusted separately by architecture and endpoint.

Architecture	Comparator	QED ≥ 0.9 difference; p_{Holm}	Mean-QED difference; p_{Holm}
RNN	Base	+2.48 [+2.31, +2.64]; 0.0059	+0.126 [+0.120, +0.133]; 0.0059
RNN	Random NE	+2.91 [+2.73, +3.09]; 0.0059	+0.142 [+0.133, +0.151]; 0.0059
RNN	Positive FT	-8.60 [-9.33, -7.87]; 0.0059	-0.078 [-0.082, -0.073]; 0.0059
Transformer	Base	-1.12 [-2.81, +0.57]; 0.1543	+0.089 [+0.055, +0.123]; 0.0078
Transformer	Random NE	+1.95 [+0.50, +3.41]; 0.0273	+0.117 [+0.053, +0.181]; 0.0078
Transformer	Positive FT	-19.19 [-21.77, -16.62]; 0.0059	-0.143 [-0.177, -0.109]; 0.0059

Table S5 Primary paired SemlaFlow comparisons for corrected negative extrapolation (CNE) at $\lambda = 2.5$. Effects are CNE minus comparator in percentage points with 95% confidence intervals. Negative liability-hit and positive 3D-usable-yield effects favor CNE. Exact two-sided sign-flip p-values are Holm-adjusted separately for each endpoint.

Comparator	Liability-hit difference; p_{Holm}	3D usable-yield difference; p_{Holm}
Base	-28.88 [-29.63, -28.13]; 0.0098	+20.81 [+19.81, +21.81]; 0.0098
Random NE	-23.48 [-25.29, -21.68]; 0.0098	+37.21 [+35.87, +38.55]; 0.0098
Positive FT	-11.11 [-11.40, -10.82]; 0.0098	+4.17 [+3.42, +4.93]; 0.0098
Standard NE	-3.09 [-3.48, -2.69]; 0.0098	+22.90 [+22.09, +23.72]; 0.0098
Norm-matched random correction	-1.21 [-1.58, -0.83]; 0.0098	+1.93 [+0.61, +3.24]; 0.0098

3 Supplementary Figures

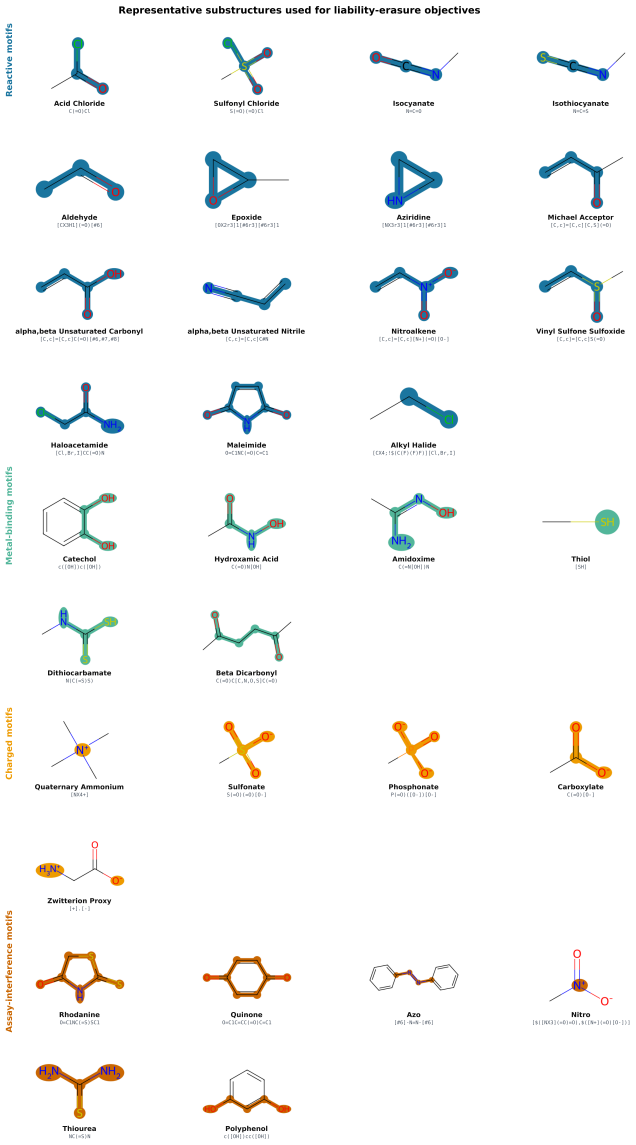


Fig. S1 Structural patterns used to define the four motif-suppression objectives. Patterns are grouped as reactive motifs, metal-binding motifs, charged motifs, and assay-interference motifs. Thick colored atoms and bonds indicate the atoms matched by the corresponding SMARTS query; thin black bonds provide molecular context in representative structures. The name and SMARTS definition printed below each structure are generated directly from the scoring implementation. A match defines membership in the operational objective but is not interpreted as proof of toxicity, assay interference, or poor developability.

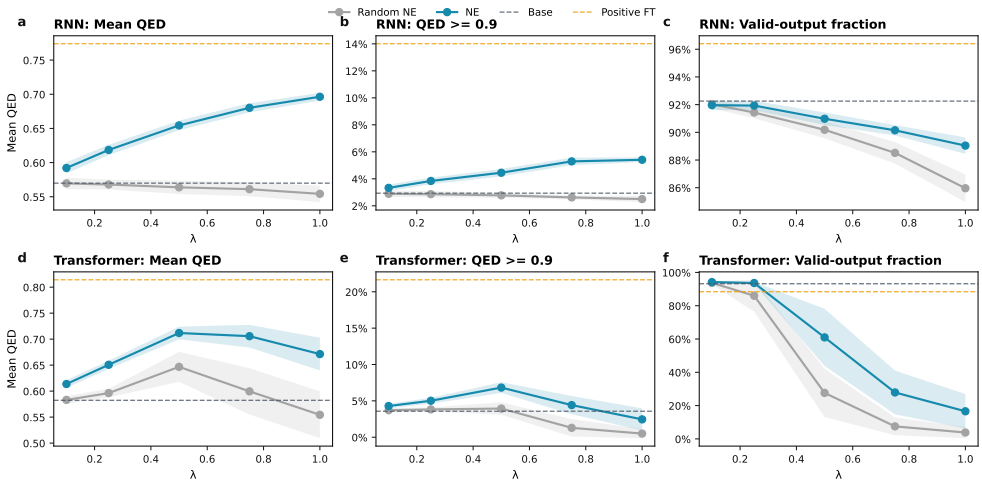


Fig. S2 Lambda-response analysis for the QED control experiment. Panels (a–c) show the RNN and panels (d–f) the Transformer. Within each row, panels report mean QED among valid outputs, the fraction of valid outputs with QED ≥ 0.9, and the valid-output fraction among all requested samples. Solid curves show means across ten paired confirmatory seeds and shaded bands show pointwise 95% Student-t confidence intervals. Dashed lines denote the base-generator and positive-fine-tuning means. These dose–response curves were treated as descriptive analyses.

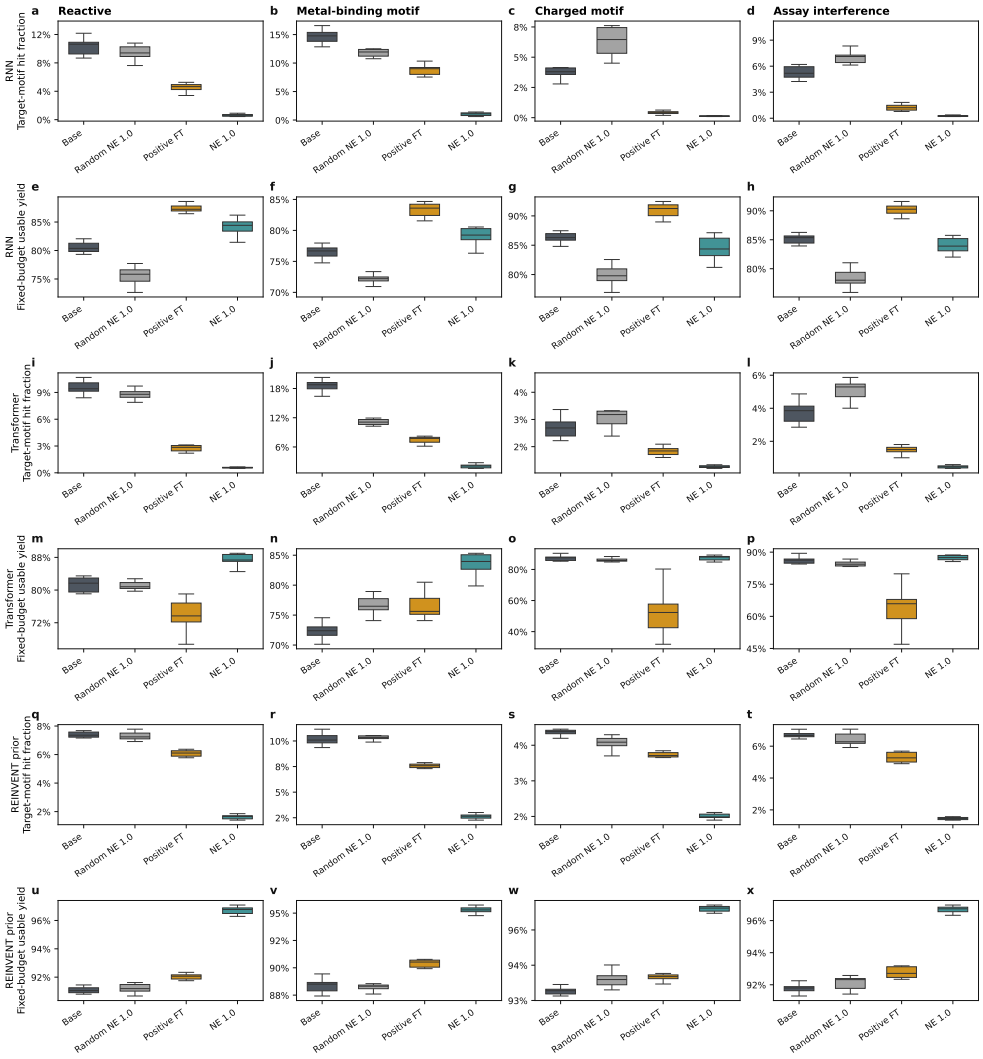


Fig. S3 Complete confirmatory endpoint results for all four motif-removal objectives across the three string-based generators. Columns correspond to reactive, metal-binding, charged, and assay-interference motifs. Panels (a–d) and (e–h) show target-motif hit fraction and fixed-budget usable yield for the RNN; panels (i–l) and (m–p) show the same endpoints for the scoped Transformer; and panels (q–t) and (u–x) show the corresponding results for the REINVENT prior. Boxes summarize ten paired confirmatory seeds; whiskers extend to 1.5 times the interquartile range. For the RNN and Transformer, usable yield requires validity, canonical uniqueness, novelty relative to the GuacaMol training set, and absence of the target motif. REINVENT-prior usable yield omits novelty because a complete canonical training reference was unavailable.

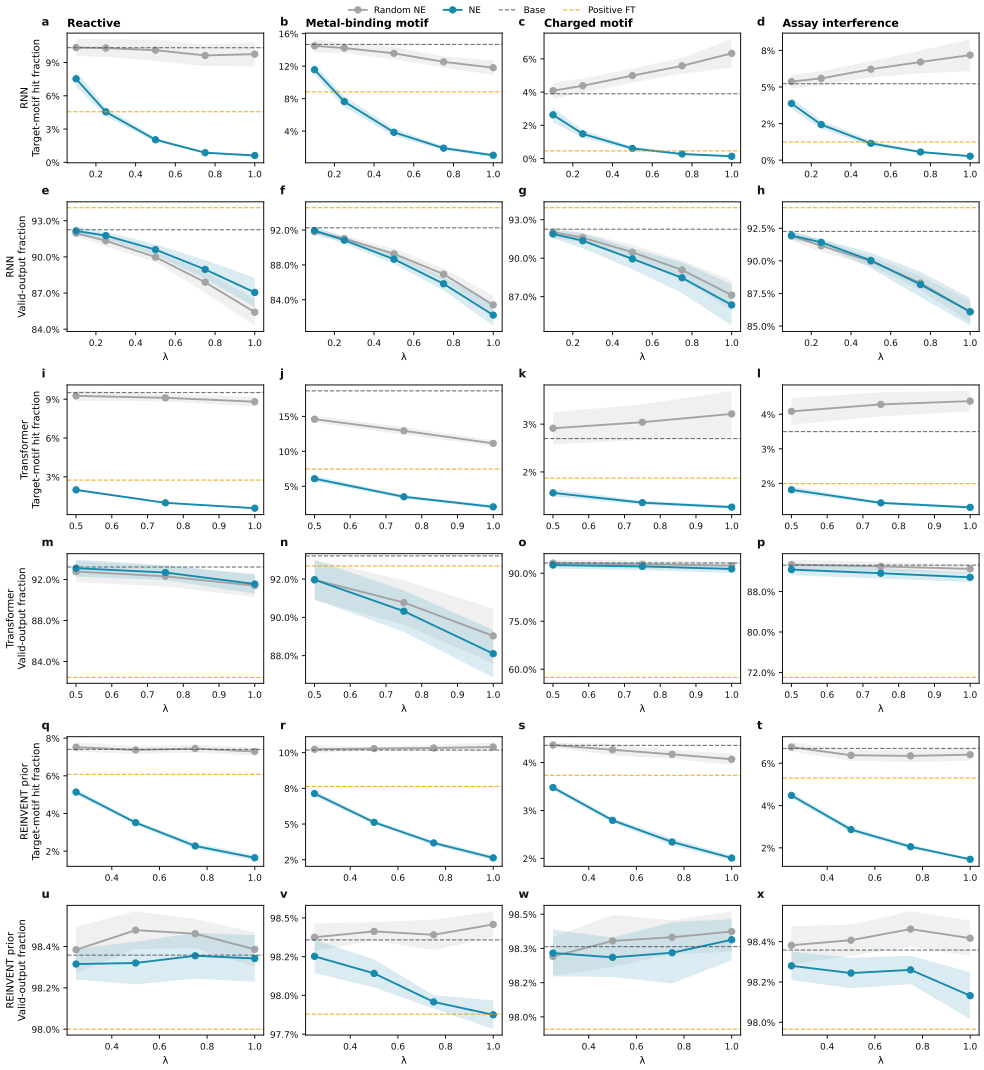


Fig. S4 Lambda responses for all four motif-removal objectives across the three string-based generators. Columns correspond to reactive, metal-binding, charged, and assay-interference motifs. Panels (a–d) and (e–h) show target-motif hit fraction and valid-output fraction for the RNN; panels (i–l) and (m–p) show the same endpoints for the scoped Transformer; and panels (q–t) and (u–x) show target-motif hit fraction and valid-output fraction for the REINVENT prior. Solid curves show means across ten paired confirmatory seeds, shaded bands show pointwise 95% Student-t confidence intervals, and dashed lines denote base-generator and positive-fine-tuning means. Each generator retains the lambda grid used in its experiment. These dose–response analyses were treated as descriptive.

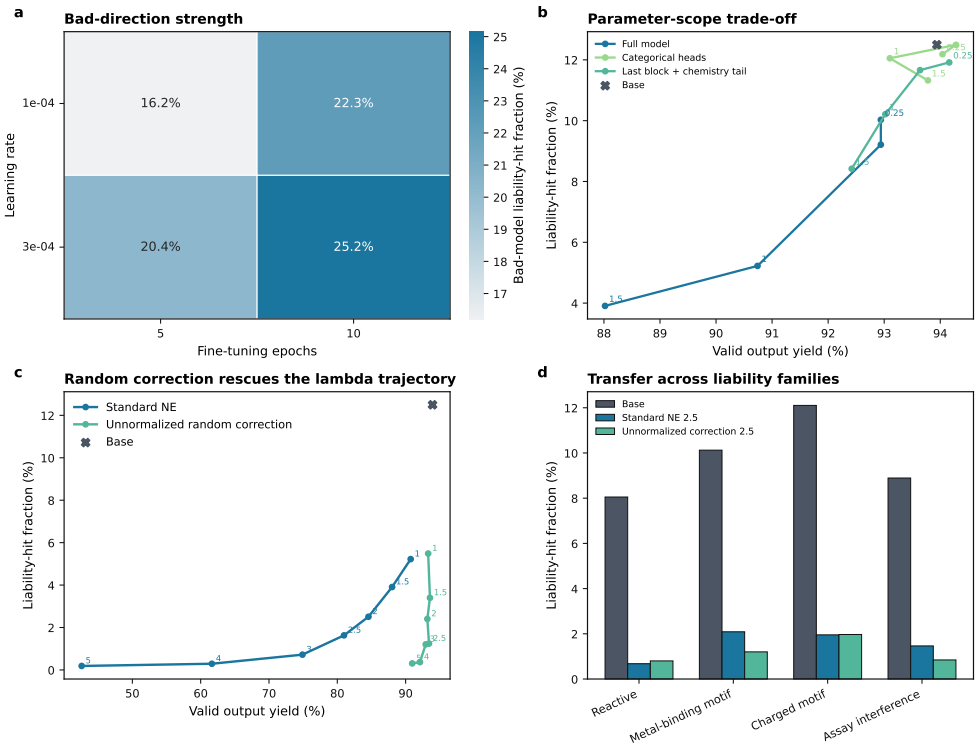


Fig. S5 Exploratory SemlaFlow parameter selection on development seed 11 using the oxygen–nitrogen single-bond proxy objective. Panel (a) shows liability frequency in negative-set fine-tuned models across learning rates and epoch counts. Panel (b) shows the trade-off between target removal and graph validity for full-model and restricted update scopes; labels denote λ. Panel (c) shows the lambda response for standard NE and the original unnormalized random-corrected diagnostic used during development. In the confirmatory experiment, the random-corrected direction was retained only as a mechanistic control and norm-matched to the CNE direction as defined in Supplementary Section 1.8. Panel (d) shows single-seed transfer of the selected operating point to the four medicinal-chemistry liability families. These panels document configuration development and are not treated as inferential replicates.

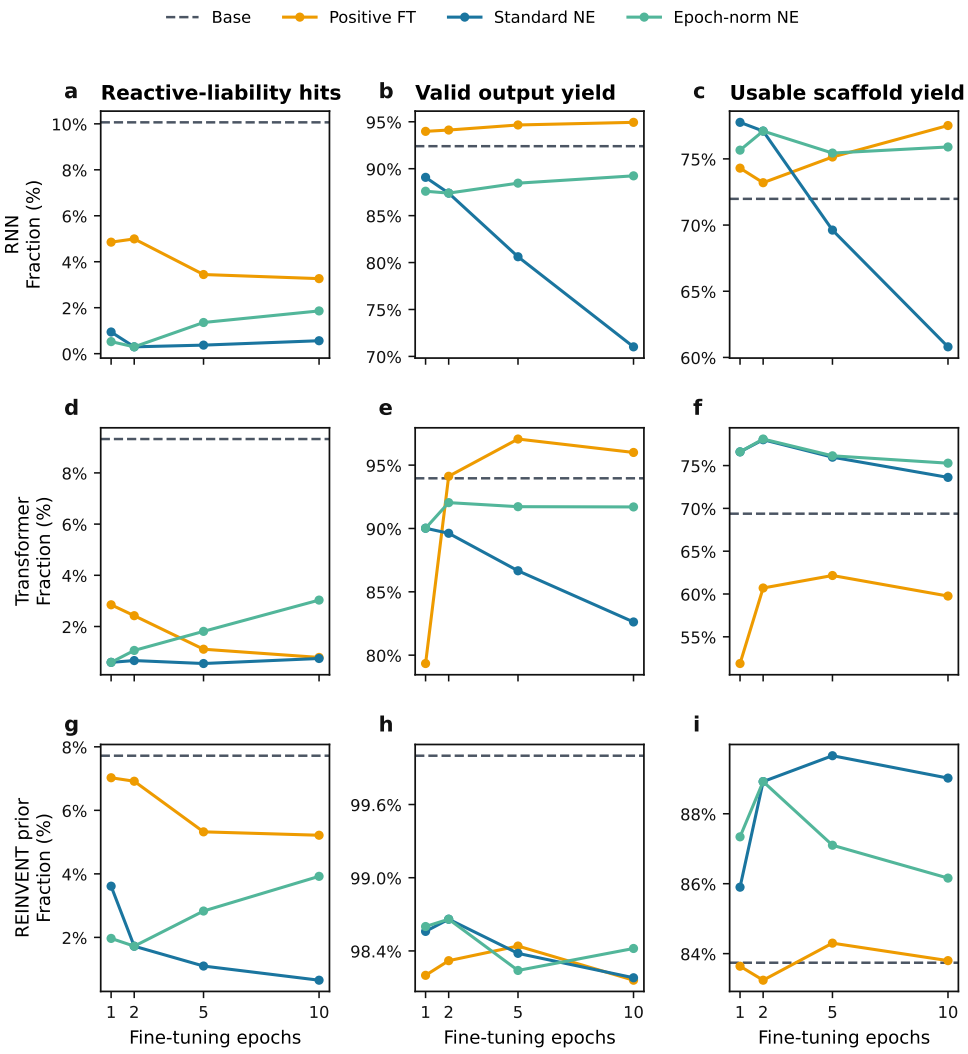


Fig. S6 Exploratory fine-tuning epoch sensitivity for reactive-liability removal on development seed 11. Panels (a–c), (d–f), and (g–i) show the RNN, Transformer, and REINVENT prior, respectively. Within each row, the left panel reports reactive-liability hit fraction, the middle panel valid-output yield relative to the fixed request of 5000 samples, and the right panel fixed-budget unique usable scaffold yield. Positive fine-tuning and negative-set fine-tuning were continued for 1, 2, 5, and 10 epochs. Standard NE uses each resulting negative-set direction directly; epoch-norm NE rescales it to the architecture's established direction norm (two epochs for the RNN, one epoch for the Transformer, and two epochs for the REINVENT prior). Dashed lines denote the base generator. These data are a single-seed sensitivity analysis and are not included in confirmatory statistical tests.

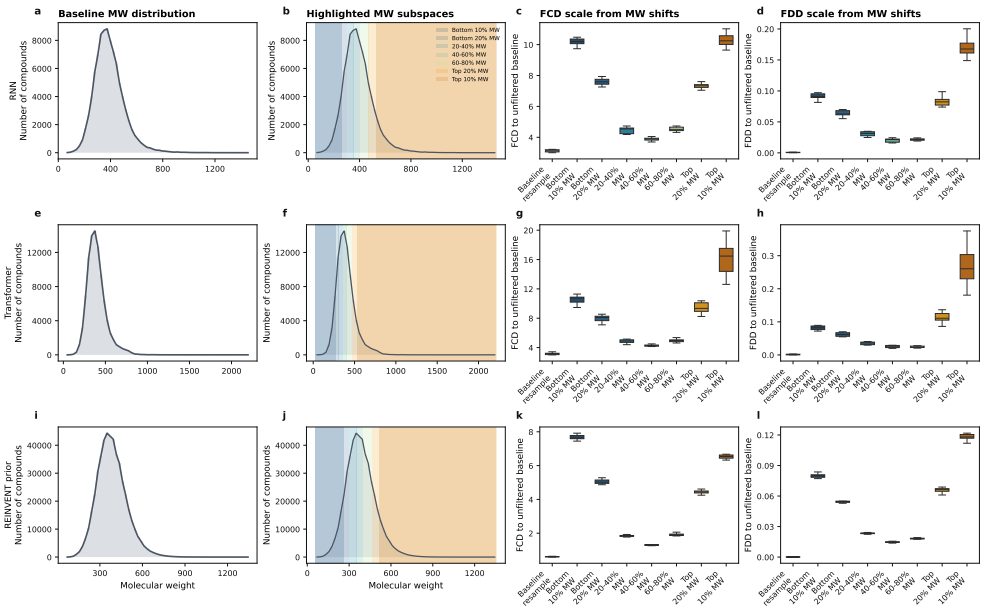


Fig. S7 Generator-specific FCD and FDD calibration using equal-sized molecular-weight subsets. Panels (a–d), (e–h), and (i–l) show the RNN, Transformer, and REINVENT prior, respectively. Within each row, the first panel shows the pooled molecular-weight distribution of valid base-generated molecules; the second highlights the bottom 10%, bottom 20%, 20–40%, 40–60%, 60–80%, top 20%, and top 10% bands. Highlights illustrate pooled quantiles; the actual query bands were defined separately for each seed. The third and fourth panels show FCD and FDD to an unfiltered draw from the same saved base pool. "Baseline resample" uses an unfiltered query. Query and reference draws may overlap. Every band and reference within each seed used the same count: 916–930 molecules for the RNN, 903–952 for the Transformer, and 4905–4916 for the REINVENT prior across seeds. Boxes summarize ten seed-level values. These controls illustrate molecular-weight restriction, not numerical thresholds for model comparisons with different sample counts.

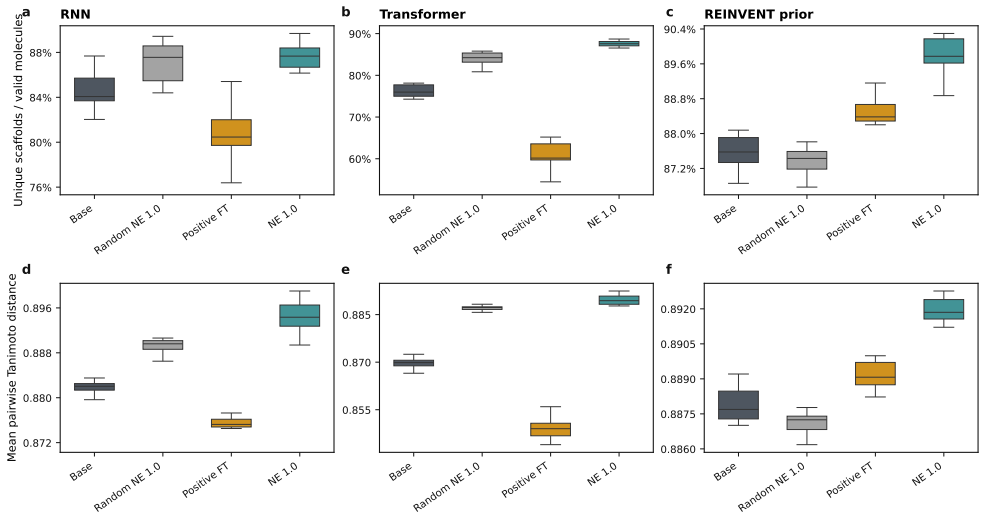


Fig. S8 Structural diversity in the representative metal-binding-motif-removal experiment. Panels (a–c) show the unique non-empty Bemis–Murcko scaffold fraction among valid generated molecules for the RNN, scoped Transformer, and REINVENT prior, respectively. Panels (d–f) show the corresponding mean pairwise Morgan-fingerprint Tanimoto distance, where larger values indicate lower average structural similarity within the generated set. Boxes summarize the ten confirmatory seeds. These analyses are descriptive.

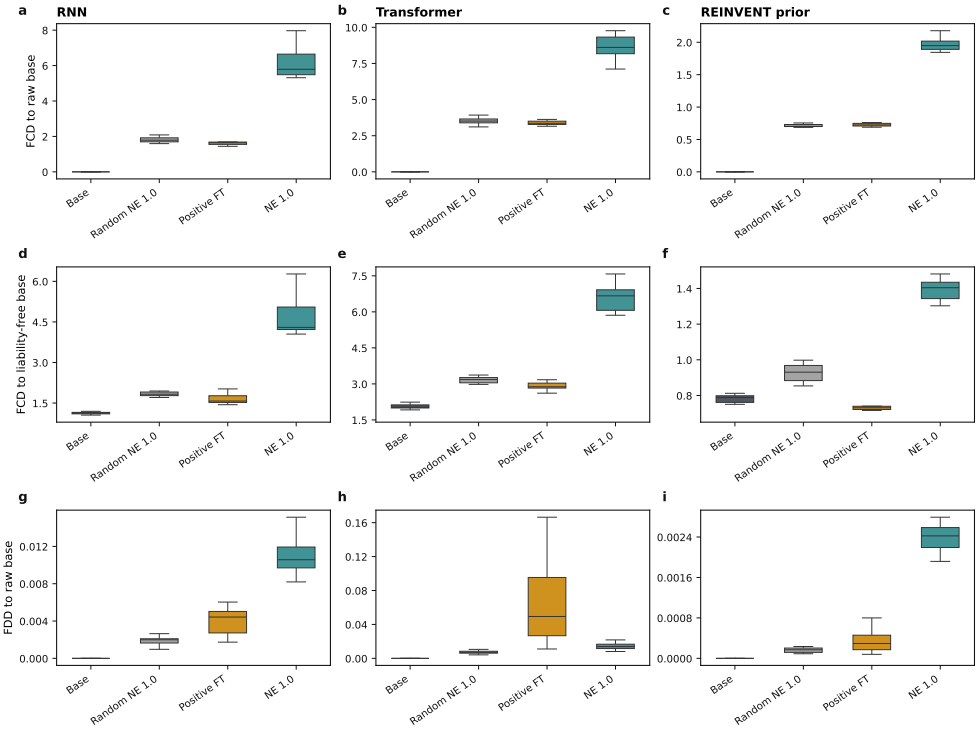


Fig. S9 Chemical-space displacement in the representative metal-binding-motif-removal experiment. Columns show the RNN, scoped Transformer, and REINVENT prior. Panels (a–c) report FCD to the saved raw-base reference, panels (d–f) FCD to the metal-binding-motif-free subset of that base reference, and panels (g–i) FDD to the raw-base reference. Base denotes the saved reference pool; its raw-base self-comparisons are zero by construction, not baseline-resampling estimates. FDD is calculated from scaled logP, molecular weight, hydrogen-bond donor count, ring count, and TPSA distributions. Boxes summarize the ten confirmatory seeds. The molecular-weight controls in Supplementary Figure S7 provide qualitative context; their sample counts differ from these comparisons.

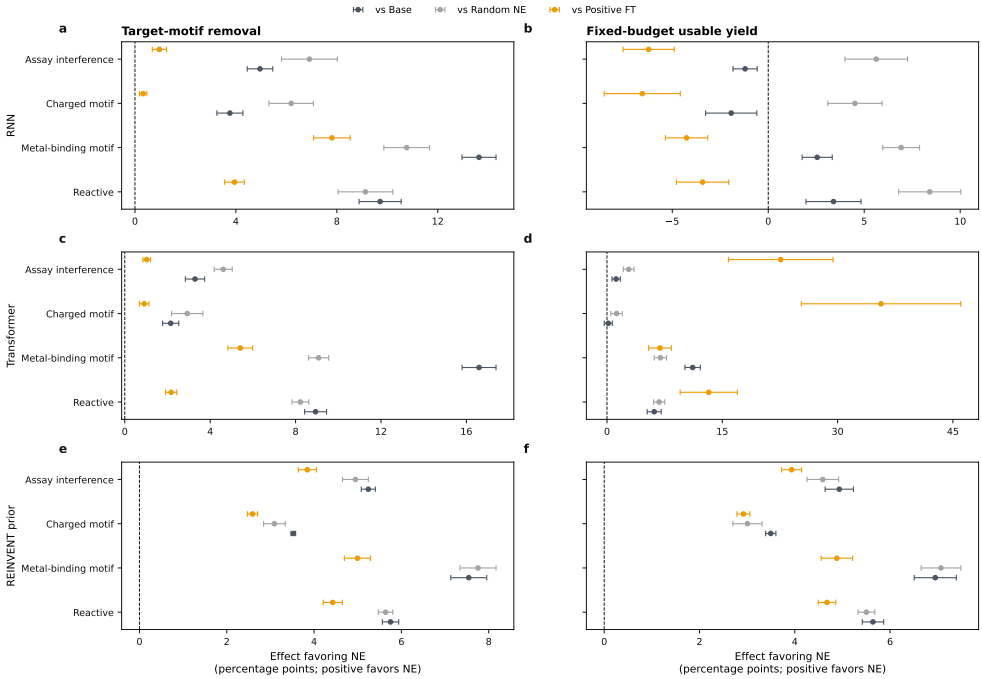


Fig. S10 Confirmatory paired effects of NE relative to base generation, norm-matched trained-random NE, and positive fine-tuning. Panels (a, c, e) show effects on target-motif removal for the RNN, scoped Transformer, and REINVENT prior, respectively. Panels (b, d, f) show effects on fixed-budget usable yield for the same generators. Target-hit differences are sign-reversed so that positive values indicate stronger motif removal by NE; usable-yield differences are plotted as NE minus comparator. Positive values therefore favor NE in every panel. Points are mean paired effects across ten confirmatory seeds and horizontal bars are two-sided 95% Student-t confidence intervals. For the RNN and Transformer, usable yield includes novelty relative to the GuacaMol training set; REINVENT-prior yield omits novelty because a complete canonical training reference was unavailable. Exact sign-flip p-values and Holm-adjusted values for the primary comparison with positive fine-tuning are reported in Supplementary Table S3.

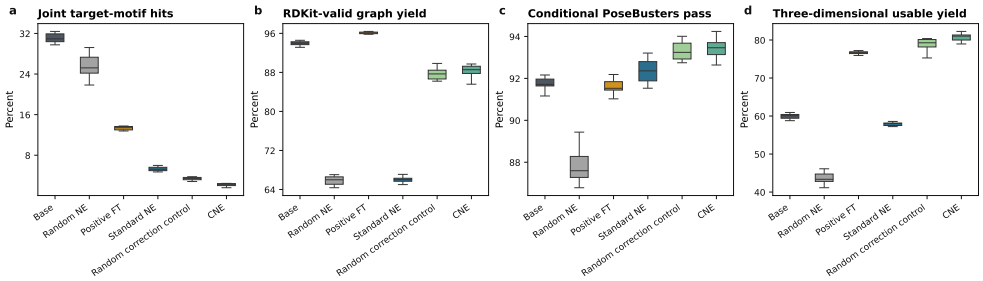


Fig. S11 SemlaFlow endpoint distributions for the base generator, controls, standard NE, and CNE in the joint motif-removal experiment. Panels show (a) joint target-motif hit fraction among RDKit-valid outputs, (b) RDKit-valid graph yield among all requested samples, (c) the conditional PoseBusters pass rate among RDKit-valid outputs, and (d) three-dimensional usable yield among all requested samples. CNE subtracts the task vector learned from liability-free molecules from the negative-set task vector. The random-correction control instead subtracts a size-matched random-set task vector and is globally norm-matched to the CNE direction. Boxes summarize ten paired confirmatory seeds; whiskers extend to 1.5 times the interquartile range.

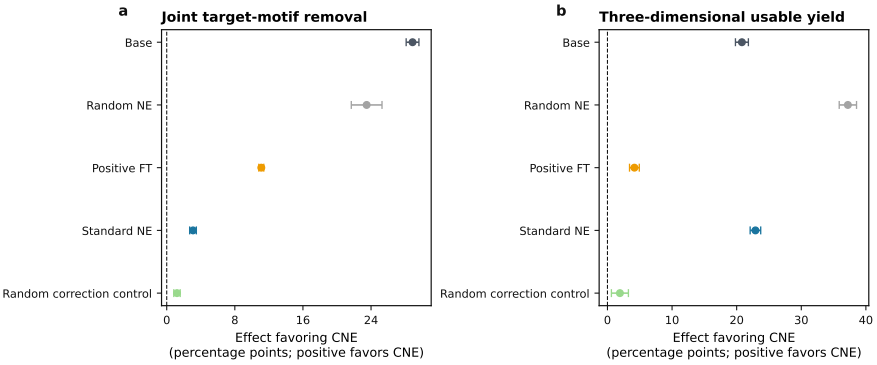


Fig. S12 Confirmatory paired effects of CNE in the SemlaFlow joint motif-removal experiment. Panel (a) shows the reduction in joint target-motif hit fraction relative to base generation, random NE, positive fine-tuning, standard NE, and the norm-matched random-correction control. Panel (b) shows the corresponding effects on three-dimensional usable yield. Target-hit differences are sign-reversed and usable-yield differences are plotted as CNE minus comparator, so positive values favor CNE in both panels. Points are mean paired effects across ten confirmatory seeds and horizontal bars are two-sided 95% Student-t confidence intervals. Exact sign-flip p-values and Holm-adjusted values are reported in Supplementary Table S5.

4 Key Reproduction Targets

The principal commands and notebooks used to reproduce the reported analyses are listed in Supplementary Table S6.

Table S6: Primary Makefile targets used to reproduce the main analyses.

Analysis	Make target and notebooks
GuacaMol base replicates	`make paper-guacamol-base-replicates`
GuacaMol liability replicates	`make paper-guacamol-liability-replicates`
GuacaMol epoch-sensitivity scaffold analysis	`make guacamol-epoch-sensitivity-scaffolds`
REINVENT epoch-sensitivity analysis	`make reinvent-reactive-epoch-sensitivity`
REINVENT liability replicates	`make paper-reinvent-liability-replicates`
SemlaFlow base replicates	`make semlaflow-geom-drugs-50000-baseline-replicates`
SemlaFlow joint-liability replicates	`make semlaflow-four-liability-cne-replicates`
SemlaFlow PoseBusters analysis	`make semlaflow-four-liability-joint-replicates-posebusters`
SemlaFlow structural analysis	`make semlaflow-four-liability-structural-analysis`
SemlaFlow endpoint/statistical analysis	`make semlaflow-four-liability-paper-analysis`
SemlaFlow usable-scaffold analysis	`make semlaflow-four-liability-usable-scaffolds`
FCD/FDD distances	`make paper-fcd-distances`
FCD/FDD MW calibration	`make paper-fcd-mw-calibration`
MW calibration with matched sample counts	`make paper-fcd-mw-calibration-size-check`
Independent REINVENT Base evaluation	`make paper-reinvent-base-budget`
Main paper figures	`notebooks/260612_main_paper_figures.ipynb`
Supplementary figures	`notebooks/260617_si_figures_and_statistics.ipynb`
Confirmatory statistics and tables	`make paper-statistics`

Supplementary SMARTS table	`make paper-si-tables`

References

[1] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv. arXiv:1912.01703 [cs.LG] (2019). https://doi.org/10.48550/arXiv.1912.01703 . http://arxiv.org/abs/1912.01703 Accessed 2026-08-10

[2] Brown, N., Fiscato, M., Segler, M.H.S., Vaucher, A.C.: GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of Chemical Information and Modeling **59**(3), 1096–1108 (2019) https://doi.org/10.1021/acs.jcim.8b00839 . Accessed 2026-08-10

[3] Cho, K., Merrienboer, B.v., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv. arXiv:1406.1078 [cs.CL] (2014). https://doi.org/10.48550/arXiv.1406.1078 . http://arxiv.org/abs/1406.1078 Accessed 2026-08-10

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. arXiv. arXiv:1706.03762 [cs.CL] (2023). https://doi.org/10.48550/arXiv.1706.03762 . http://arxiv.org/abs/1706.03762 Accessed 2026-08-10

[5] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv. arXiv:1711.05101 [cs.LG] (2019). https://doi.org/10.48550/arXiv.1711.05101 . http://arxiv.org/abs/1711.05101 Accessed 2026-08-10

[6] Brenk, R., Schipani, A., James, D., Krasowski, A., Gilbert, I.H., Frearson, J., Wyatt, P.G.: Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. ChemMedChem **3**(3), 435–444 (2008) https://doi.org/10.1002/cmdc.200700139

[7] Gehringer, M., Laufer, S.A.: Emerging and re-emerging warheads for targeted covalent inhibitors: Applications in medicinal chemistry and chemical biology. Journal of Medicinal Chemistry **62**(12), 5673–5724 (2019) https://doi.org/10.1021/acs.jmedchem.8b01153

[8] Martin, D.P., Blachly, P.G., McCammon, J.A., Cohen, S.M.: Exploring the Influence of the Protein Environment on Metal-Binding Pharmacophores. Journal of Medicinal Chemistry **57**(16), 7126–7135 (2014) https://doi.org/10.1021/jm500984b . Accessed 2026-08-10

[9] Baell, J.B., Holloway, G.A.: New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. Journal of Medicinal Chemistry **53**(7), 2719–2740 (2010) https://

doi.org/10.1021/jm901137j

[10] Dahlin, J.L., Nissink, J.W.M., Strasser, J.M., Francis, S., Higgins, L., Zhou, H., Zhang, Z., Walters, M.A.: PAINS in the assay: Chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS. Journal of Medicinal Chemistry **58**(5), 2091–2113 (2015) https://doi.org/10.1021/jm5019093

[11] Baell, J.B., Nissink, J.W.M.: Seven Year Itch: Pan-Assay Interference Compounds (PAINS) in 2017-Utility and Limitations. ACS Chemical Biology **13**(1), 36–44 (2017) https://doi.org/10.1021/acschembio.7b00903 . Accessed 2026-08-10

[12] Irwin, R., Tibo, A., Janet, J.P., Olsson, S.: Semlaflow – efficient 3D molecular generation with latent attention and equivariant flow matching. In: Proceedings of the 28th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 258, pp. 3772–3780. PMLR, Brookline, MA (2025). https://proceedings.mlr.press/v258/irwin25a.html

[13] Özçelik, R., Grisoni, F.: How evaluation choices distort the outcome of generative drug discovery. Journal of Cheminformatics **17**(1), 169 (2025) https://doi.org/10.1186/s13321-025-01108-y