Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Negative Extrapolation for Targeted Motif Suppression in Molecular Generators

Joshua Hesse
Technical University of Munich
Julia Westermayr
Leibniz Institute of Surface Engineering
Stephan A. Sieber
stephan.sieber@tum.de

Technical University of Munich

Research Article

Keywords:

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11270647/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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;
Contributing authors: joshua.hesse@tum.de;

Abstract

Broadly trained molecular generators can cover diverse chemical space, but may also produce motifs that are undesirable in a given discovery context, e.g., reactive motifs. Adapting a pretrained generator to suppress such chemistry is therefore an important practical problem. However, most model steering methods enrich desirable compounds or optimize a predefined reward, whereas selectively suppressing an undesirable region of chemical space without compromising the quality and diversity of the remaining distribution remains challenging. Here, we adapt negative extrapolation (NE), a parameter-space editing strategy in which a generator is fine-tuned on molecules containing unwanted structural motifs and the resulting weight update is reversed, to suppress these motifs directly within the generator. We evaluated NE using two custom string-based autoregressive generators, a public pretrained recurrent prior, and a three-dimensional flow-matching generator, with unedited generation and conventional fine-tuning

on molecules without the unwanted motifs as principal comparators. Across the string-based generators, NE reduced the occurrence of four structural motif classes (reactive, metal-binding, charged, and assay-interference motifs) more strongly than conventional fine-tuning. We further extended NE through selective parameter editing and combinations of learned update directions, enabling generative validity to be preserved across different model architectures. In the three-dimensional flow-matching generator, corrected NE reduced the fraction of valid molecules containing any of the unwanted motifs from 31.1% to 2.2% and increased the fixed-budget yield of unique, motif-free molecules that were valid molecular graphs and passed three-dimensional chemical and geometric consistency checks from 60.0% to 80.8%. Together, these results support NE as a simple approach for selectively editing molecular generators across architectures and across string-based and three-dimensional representations, while preserving molecular diversity and without requiring a reward-optimization loop.

Scientific Contribution
We apply the parameter-arithmetic principle of reversing a learned update to an important problem in molecular generation: selectively suppressing a defined unwanted region of chemical space while preserving the quality and diversity of the remaining distribution. Across string-based and three-dimensional generators, we further investigate how parameter arithmetic can be adapted through selective parameter editing and combinations of independently learned update directions to retain model validity when full-model extrapolation is disruptive.

Keywords: molecular generation, negative extrapolation, task arithmetic, motif suppression, three-dimensional generation, flow matching, cheminformatics

List of abbreviations

3D three-dimensional

CI confidence interval

CNE corrected NE

FCD Fréchet ChemNet distance

FDD Fréchet descriptor distance

FT fine-tuning

LBDD ligand-based drug design

LSTM long short-term memory

MGM molecular generative model

NE negative extrapolation

NEON Negative Extrapolation from Self-Training

ODE ordinary differential equation

QED quantitative estimate of drug-likeness

RNN recurrent neural network

SBDD structure-based drug design

SMARTS SMILES arbitrary target specification

SMILES simplified molecular-input line-entry system

TPSA topological polar surface area

1 Introduction

Molecular generative models (MGMs) are widely used in contemporary drug discovery, enabling rapid *in silico* exploration of chemical spaces for virtual screening and molecular design campaigns [1–4]. In typical workflows, generated molecules are filtered, scored, and prioritized according to task-relevant properties, such as drug-likeness, synthetic accessibility, and predicted activity, yielding candidate compounds for further computational or experimental evaluation [5]. Molecular generation includes both unconditional models, which aim to learn broad molecular distributions, and conditional or goal-directed models, which bias generation toward specific design criteria, for example by incorporating target-protein structure in structure-based drug design (SBDD) or known ligand information in ligand-based drug design (LBDD) [6–10]. Although practical drug discovery applications often require such conditioning, unconditional models remain an important foundation and controlled setting for method development, since they can serve as starting points for subsequent fine-tuning on task-specific data, model editing, or goal-directed optimization [11].

In addition to enriching desirable properties, medicinal chemistry workflows must exclude structures that are unsuitable for a particular application. Generated molecules may, for example, need to avoid reactive electrophiles, metal-binding groups,

structural alerts, or assay-interference motifs [12–18]. Such constraints can often be defined precisely through the presence of specific chemical motifs. Their absence, however, does not define an equally specific positive class: molecules without a given motif may occupy many otherwise unrelated regions of chemical space. Fine-tuning on these molecules is possible, but it reinforces the chemistry represented in the selected positive set rather than directly specifying which part of the learned distribution should be suppressed.

Several existing strategies can suppress such motifs. Post-generation filtering is straightforward, but discards molecules from a fixed sampling budget rather than changing the generator itself [12]. For example, in a recent structure-based molecular-generation study, only 30,662 of 69,867 chemically valid generated molecules remained after applying REOS/Dundee structural-alert filters, corresponding to the rejection of 56.1% of the valid output [19]. Conventional fine-tuning on motif-free molecules can reduce their frequency, but its effect depends on the composition of the selected set. More generally, prolonged or iterative biasing can reduce novelty and diversity, increase synthetic complexity, and introduce improbable structural arrangements [8, 20–22]. Removing motif-containing molecules from the training data and retraining the generator offers a more comprehensive solution, but becomes costly for large models or when many objectives must be considered. Reward-based methods, including REINVENT4, a reinforcement-learning framework for de novo molecular design, can directly penalize unwanted motifs through configurable scoring functions [20]. However, they require iterative sampling and optimization, and their behavior depends on how the reward is constructed and weighted [23].

Parameter-space model editing offers a different route to steering generative models. The parameter change produced by fine-tuning can be treated as a task vector and added to or subtracted from a base model to modify its behavior [24, 25]. Related approaches in image generation have used model editing to suppress named visual

concepts while preserving unrelated behavior [26, 27]. The Negative Extrapolation from Self-Training (NEON) framework introduced a simple weight-space operation in which a generator is fine-tuned on its own synthetic output and the resulting update is reversed to move away from the degradation caused by self-training [28].

In molecular generation, Özçelik and Grisoni used molecular task arithmetic to optimize specific design objectives by learning an update direction from negative examples and reversing it to enrich the generated output for the corresponding desirable property [29]. Here, we use the same general parameter-arithmetic principle for a different objective: selectively suppressing a defined unwanted region of chemical space while preserving the quality and diversity of the remaining distribution. We further investigate how such edits behave across different generator architectures, and whether restricting the edited parameter set or combining independently learned update directions can preserve generative quality when full-model extrapolation is disruptive.

To test this, we apply NE to the selective suppression of predefined unwanted motifs from unconditional molecular generators. Fine-tuning on motif-containing molecules estimates an unwanted parameter direction, and extrapolating from the base model in the opposite direction produces the edited generator. We evaluate this strategy in recurrent and Transformer models, a public pretrained REINVENT prior distributed with REINVENT4 [20], and the three-dimensional flow-matching model SemlaFlow [30]. The latter extends the evaluation beyond string-based generation to a model that jointly produces molecular graphs and three-dimensional (3D) conformations. Across these systems, we assess motif suppression, molecular validity, three-dimensional quality, chemical-space coverage, structural diversity, and the yield of usable molecules under a fixed sampling budget.

2 Study overview and model editing

2.1 Study design and workflow

All experiments followed the workflow in Figure 1. A base generator produced an initial molecular pool, from which molecules satisfying the objective-specific negative criterion were selected (panels 1–3). For motif-suppression experiments, motif-free molecules provided the positive fine-tuning comparator, while a size-matched random subset provided the control. Separate copies of the same base checkpoint were fine-tuned on these sets using the model's native generative objective, shown for the negative set in panels 4,5. The negative-set update was then reversed to construct the NE model (panels 6,7). No binary labels, classifier loss, or reward function was introduced.

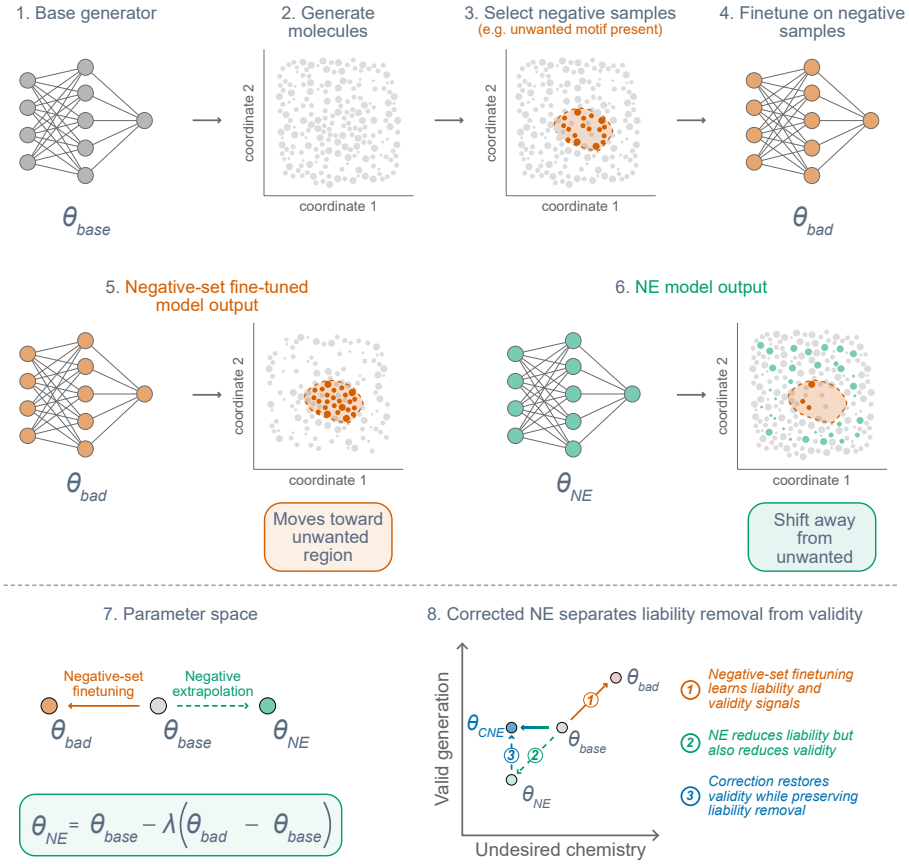


Fig. 1 Conceptual workflow for negative extrapolation (NE) and corrected NE (CNE). (1–3) A base generator is sampled and molecules containing a predefined unwanted motif are selected as the negative set. (4,5) Fine-tuning on this set produces θ_{bad} and shifts the generated distribution toward the unwanted region. (6,7) NE reverses the negative-set update according to $\theta_{\text{NE}} = \theta_{\text{base}} - \lambda(\theta_{\text{bad}} - \theta_{\text{base}})$, thereby shifting generation away from that region. (8) Conceptual rationale for CNE: the negative-set update may contain changes associated with both the unwanted motif and valid molecular generation. Reversing the complete update can therefore reduce validity alongside the motif. CNE combines the negative-set update with a correction learned from motif-free molecules, aiming to restore validity while retaining motif removal. The chemical-space distributions and parameter-space directions are schematic.

We first tested quantitative estimate of drug-likeness (QED) steering as a general quality-control objective in a recurrent neural network (RNN) and a Transformer trained here on GuacaMol, a dataset of molecular simplified molecular-input line-entry system (SMILES) derived from ChEMBL 24 [31, 32]. We then evaluated suppression of four RDKit-defined motif classes (reactive, metal-binding, charged,

and assay-interference motifs) in these models and the public REINVENT prior, a ChEMBL-pretrained recurrent SMILES generator distributed with REINVENT4 [20]. These three models are referred to as the string-based generators. Finally, we targeted molecules containing any of the four motif classes in SemlaFlow, an equivariant flow-matching generator pretrained on GEOM-Drugs that jointly produces molecular graphs and three-dimensional conformations [30, 33]. Table 1 summarizes the generators and their training sources. We also examined restricted parameter updates and correction directions to preserve generative validity.

Evaluation used new fixed-budget samples to assess the target response alongside molecular validity, structural diversity, chemical-space coverage, and usable yield (Equation 6). For SemlaFlow, this additionally included PoseBusters, a suite of chemical and geometric consistency checks for generated molecular structures [34]. Confirmatory comparisons used ten paired seeds and were kept separate from exploratory method-development experiments. Sampling budgets, objective definitions, generator-specific acceptance criteria, and statistical procedures are detailed in the Methods (Section 6).

Table 1 Overview of the molecular generators, their representations, architectures, and training sources.

Generator	Representation and architecture	Base model and training source
RNN	SMILES; three-layer GRU	Trained in this study for eight epochs on 1,273,104 GuacaMol training SMILES derived from ChEMBL 24
Transformer	SMILES; four causal Transformer blocks	Trained in this study for eight epochs on the same GuacaMol set
REINVENT prior	SMILES; three-layer LSTM	Public ChEMBL-trained prior distributed with REINVENT4
SemlaFlow	Molecular graph and three-dimensional coordinates; 12-layer equivariant flow-matching model	Public 200-epoch GEOM-Drugs checkpoint

2.2 Negative extrapolation and control directions

Let θ_{base} denote the parameters of a base molecular generator. For a given objective, molecules are sampled from the base generator and a negative set is selected, for example by retaining molecules that match a targeted structural-motif flag. Fine-tuning on this set with the architecture's native generative training objective produces a negative-set fine-tuned checkpoint θ_{bad}. NE constructs

$$\theta_{\text{NE}}(\lambda) = \theta_{\text{base}} - \lambda(\theta_{\text{bad}} - \theta_{\text{base}}), \tag{1}$$

where λ controls the edit strength. Algebraically, this operation is a form of task-vector negation [24, 25, 29]. In our setting, the vector is obtained by continuing generative training on a model-generated subset of molecules matching the unwanted motif, rather than by supervised fine-tuning on a separate labelled task.

For motif-removal objectives, fine-tuning on molecules that did not contain the unwanted motif analogously produced the motif-free checkpoint θ_{free}. Unless otherwise stated, the motif-containing and motif-free sets were balanced by downsampling the larger set.

We denote the corresponding task vectors by $\Delta_{\text{bad}} = \theta_{\text{bad}} - \theta_{\text{base}}$ and $\Delta_{\text{free}} = \theta_{\text{free}} - \theta_{\text{base}}$. As illustrated in Figure 1, panel 8, fine-tuning on valid molecules containing an unwanted motif may encode two effects in the same task vector: increased generation of the selected motif and more general changes that support valid molecular generation. Reversing the complete update can therefore suppress the unwanted motif while also weakening molecular validity. Corrected NE (CNE) attempts to separate these effects by using an update learned from a broad set of motif-free valid molecules to estimate and cancel the shared valid-generation component before the remaining motif-associated direction is reversed. We therefore hypothesize the following first-order decomposition, which was not independently verified at the weight

level:

$$\begin{aligned} \Delta_{\text{bad}} &\approx \Delta_{\text{shared}} + \Delta_{\text{liability}}, \\ \Delta_{\text{free}} &\approx \Delta_{\text{shared}}. \end{aligned} \tag{2}$$

Here, Δ_{shared} represents changes associated with continued training on valid generated molecules, while $\Delta_{\text{liability}}$ represents changes associated with the selected unwanted chemistry. The second line treats the broad motif-free set primarily as an estimate of the shared generative component; it does not require the update to be exactly free of smaller subset-specific effects. Under this hypothesis, subtracting the motif-free update from the negative-set update cancels the changes shared by both directions:

$$\Delta_{\text{bad}} - \Delta_{\text{free}} \approx \Delta_{\text{liability}}. \tag{3}$$

For SemlaFlow, we formed CNE by subtracting the independently learned motif-free task vector before extrapolation:

$$\theta_{\text{CNE}}(\lambda) = \theta_{\text{base}} - \lambda(\Delta_{\text{bad}} - \Delta_{\text{free}}). \tag{4}$$

The motif-free direction is used here as a validity correction rather than as a separate enrichment objective. Under the proposed decomposition, $\Delta_{\text{bad}} - \Delta_{\text{free}}$ provides a more selective estimate of the motif-associated task direction than Δ_{bad} alone. All task vectors were estimated independently from the same base checkpoint using the unchanged generative training objective; no negative labels or contrastive loss were introduced.

The main comparative baselines were positive fine-tuning and random NE. Random NE was constructed by fine-tuning a base-model copy on a size-matched random

molecular set, producing $\Delta_{\text{rand}} = \theta_{\text{rand}} - \theta_{\text{base}}$. This task vector was globally norm-matched to the negative-set update using

$$s_{\text{rand}} = \frac{\|\Delta_{\text{bad}}\|_2}{\|\Delta_{\text{rand}}\|_2}, \qquad \theta_{\text{random NE}}(\lambda) = \theta_{\text{base}} - \lambda s_{\text{rand}} \Delta_{\text{rand}}. \tag{5}$$

Random NE therefore tests whether an arbitrary trained direction can reproduce the effect at a matched parameter displacement.

The evaluated update configurations and their roles as methods, comparators, or controls are summarized in Table 2.

Table 2 Summary of the model updates evaluated in the motif-removal experiments. All task vectors were learned independently from the same base checkpoint using the model's native generative training objective.

Configuration	Applied parameters	Fine-tuning data	Purpose
Positive fine-tuning	$\theta_{\text{free}} = \theta_{\text{base}} + \Delta_{\text{free}}$	Molecules without the target motif	Comparator: conventional fine-tuning toward motif-free chemistry.
NE	$\theta_{\text{base}} - \lambda\Delta_{\text{bad}}$	Molecules containing the target motif	Method: reverses the update learned from the unwanted molecular subset.
Random NE	$\theta_{\text{base}} - \lambda s_{\text{rand}}\Delta_{\text{rand}}$	Size-matched random molecules	Control: tests whether an arbitrary trained direction with the same parameter-space norm reproduces the NE effect.
CNE	$\theta_{\text{base}} - \lambda(\Delta_{\text{bad}} - \Delta_{\text{free}})$	Target-motif and motif-free sets	Corrected method: aims to retain motif removal while canceling shared changes associated with valid molecular generation.

Task vectors are defined as $\Delta_{\text{bad}} = \theta_{\text{bad}} - \theta_{\text{base}}$, $\Delta_{\text{rand}} = \theta_{\text{rand}} - \theta_{\text{base}}$, and $\Delta_{\text{free}} = \theta_{\text{free}} - \theta_{\text{base}}$. Here, $s_{\text{rand}} = \|\Delta_{\text{bad}}\|_2 / \|\Delta_{\text{rand}}\|_2$.

3 Results

3.1 Positive fine-tuning outperforms NE when high-scoring examples are abundant

As a first task, we asked whether moving away from low-scoring molecules through NE could improve molecular outputs as effectively as positive fine-tuning when high-scoring examples were readily available. Motivated by NEON's quality-improvement experiments, we selected QED, a composite measure of drug-likeness, as a practical molecular optimization target [35].

For each seed, valid molecules sampled from the RNN and Transformer base models were ranked by QED; the bottom 10% formed the negative set and the top 10% formed the positive fine-tuning set. At $\lambda = 1.0$, NE increased mean QED from 0.570 to 0.696 for the RNN and from 0.582 to 0.671 for the Transformer (Figure 2a,b). Across the tested lambda range, increasing the extrapolation strength progressively increased QED, but also reduced the valid-output fraction, particularly for the Transformer (Supplementary Figure S2). The initial samples generated for each model and seed also contained high-QED molecules that could be selected directly for positive fine-tuning. Positive fine-tuning reached mean QED values of 0.774 for the RNN and 0.814 for the Transformer. Consistent with the mean-QED results, NE also increased the fraction of high-QED outputs (QED ≥ 0.9) for the RNN, but produced little improvement for the Transformer at $\lambda = 1.0$ (Figure 2c,d). Positive fine-tuning performed substantially better, exceeding NE by 8.60 percentage points for the RNN (95% CI, 7.87–9.33) and by 19.19 percentage points for the Transformer (95% CI, 16.62–21.77). Both differences remained significant after Holm correction (Supplementary Table S4).

Improving the target score is only useful if the generator continues to produce valid molecules. At $\lambda = 1.0$, the valid-output fraction of the RNN decreased from 92.3% to 89.0%, whereas full-model Transformer NE reduced the valid SMILES fraction

from 93.2% to 16.6% (Figure 2e,f). Taken together, these results show that NE can improve the generated QED distribution, but that positive fine-tuning is more effective when strong positive examples are readily available. This difference was particularly apparent in the fraction of outputs with QED ≥ 0.9. Suppressing low-scoring chemistry does not explicitly favor the highest-scoring molecules, whereas positive fine-tuning learns directly from them. These findings motivated our subsequent focus on structural motifs, where the unwanted chemistry is clearly defined but its absence leaves a broad, chemically heterogeneous set rather than a distinct positive class. Full-model NE also caused severe validity loss in the Transformer, prompting us to investigate whether restricting the edited parameters could preserve generative validity.

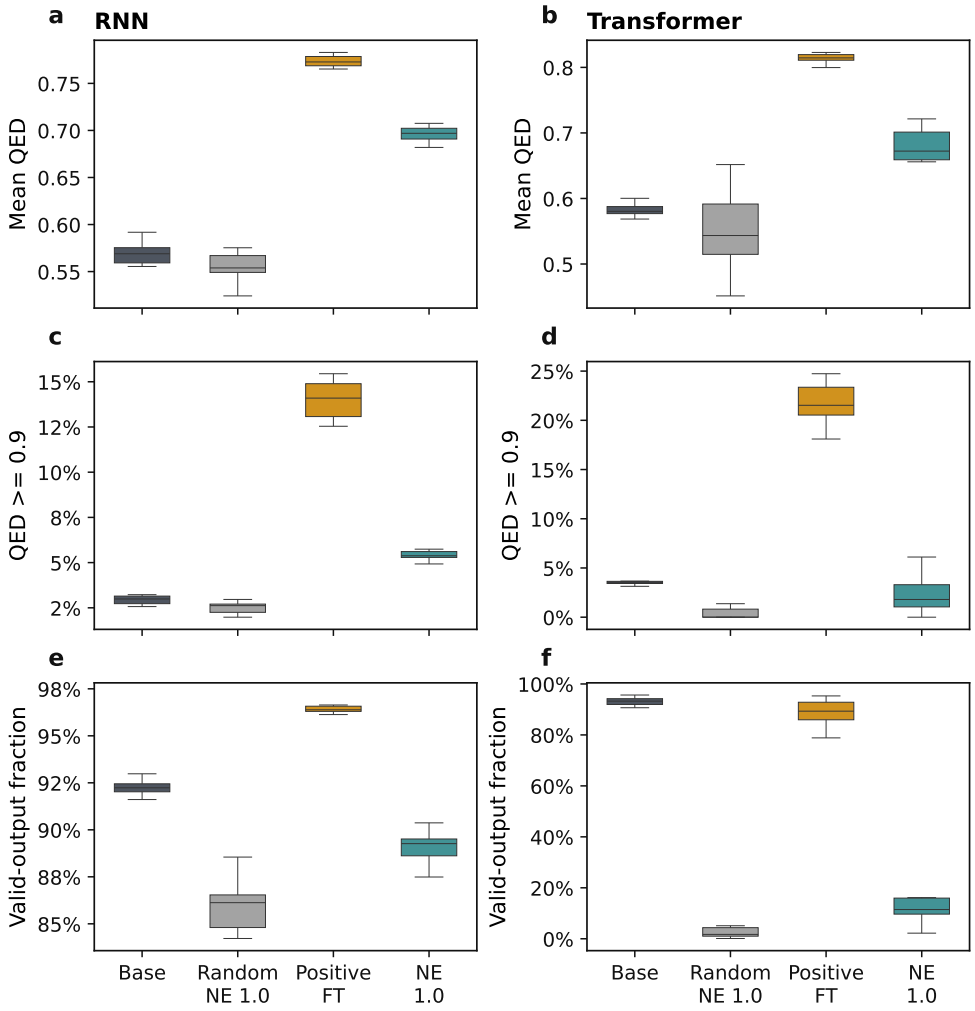


Fig. 2 QED control experiment across ten paired confirmatory seeds. Columns show the RNN (a,c,e) and Transformer (b,d,f). Rows report mean QED among valid outputs (a,b), the fraction of valid outputs with QED ≥ 0.9 (c,d), and the valid-output fraction among all requested samples (e,f) for base generation, norm-matched trained-random NE, positive fine-tuning, and NE at $\lambda = 1.0$. Boxes summarize the ten seed-level values; whiskers extend to 1.5 times the interquartile range. Lambda-response curves are reported in Supplementary Figure S2, and paired comparisons in Supplementary Table S4.

3.2 Scoped Transformer updates mitigate validity collapse

The QED experiment showed that applying NE to all Transformer parameters could severely impair molecular validity. This would limit the use of NE in architectures that are sensitive to broad parameter changes. We therefore investigated whether modifying only part of the network could retain the target-removal effect while preserving valid generation. Before applying NE in the confirmatory Transformer motif-removal experiments, we compared several restricted parameter updates using metal-binding-motif removal as a representative objective.

Five update scopes were evaluated across five λ values using development seeds 5, 7, and 11. At $\lambda = 1.0$, full-model NE reduced the mean metal-binding-motif hit fraction from 19.1% at baseline to 1.7%, but reduced validity from 94.4% to 7.8%. Restricting the update to the final Transformer block and output layer retained nearly the same liability response, reaching a 2.2% hit fraction, while preserving 88.1% validity. Updating only the output layer preserved 93.4% validity but left a substantially higher hit fraction of 9.5% (Figure 3a,b). The corresponding motif-hit and validity responses across the tested lambda values are shown in Figure 3c,d. The final-block-plus-output update provided the clearest compromise between target removal and generative validity and was therefore used for the subsequent confirmatory Transformer experiments.

These results show that the Transformer validity collapse was not an unavoidable consequence of NE, but depended strongly on which parameters were modified. More generally, this experiment demonstrates that adapting the edited parameter set to the model architecture can retain strong motif removal while avoiding the severe validity loss caused by unrestricted full-model extrapolation.

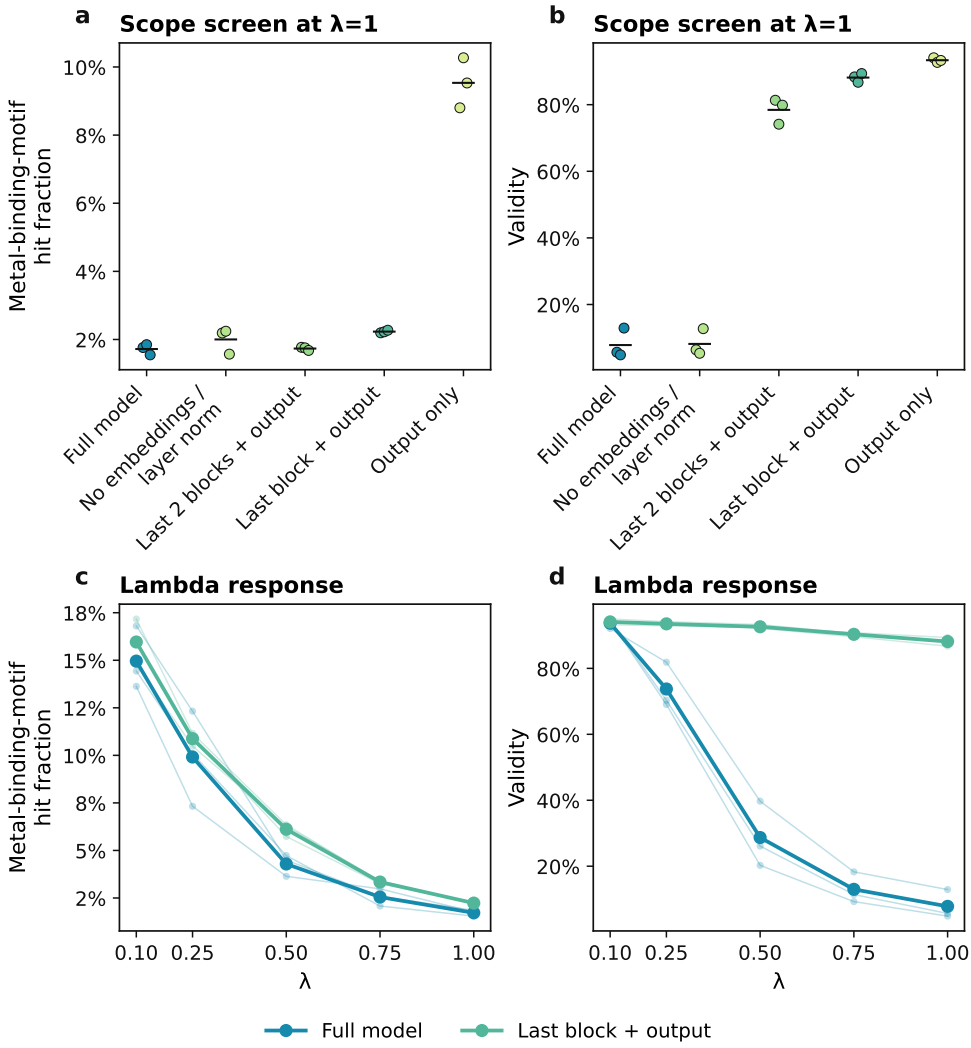


Fig. 3 Exploratory Transformer scope ablation for metal-binding-motif removal using development seeds 5, 7, and 11. Panels (a) and (b) compare the metal-binding-motif hit fraction and valid-output fraction, respectively, at $\lambda = 1.0$ for full-model NE and four restricted parameter scopes. Points show individual development seeds and horizontal bars show their means. Panels (c) and (d) show the corresponding hit-fraction and validity responses across $\lambda \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$ for full-model NE and the selected final-block-plus-output scope. Thin lines show individual seed trajectories and thick lines connect pointwise means. Every scope–lambda combination was evaluated with the same three seeds.

3.3 Targeted motif suppression transfers across string-based generators

The QED control represents an objective for which favorable examples are abundant and readily selected, making positive fine-tuning particularly effective. We therefore next considered the complementary setting: the undesirable chemistry is defined by a specific structural motif, whereas the acceptable set is the broad and chemically heterogeneous complement. In this setting, NE can learn directly from molecules that express the feature to be suppressed, rather than asking positive fine-tuning to represent all possible liability-free chemistry.

We evaluated four such structural classes—reactive motifs, metal-binding motifs, charged motifs, and assay-interference motifs—in the RNN, Transformer, and public pretrained REINVENT prior. Figure 4 shows the metal-binding objective as a matched representative example across the three generators, with target-hit fractions in panels a–c; results for all four objectives are reported in Supplementary Figure S3, and the corresponding lambda responses in Supplementary Figure S4. Across all twelve architecture–objective combinations, NE reduced the target feature more strongly than positive fine-tuning and norm-matched random NE. Relative to positive fine-tuning, target-hit fractions were 0.32–7.80 percentage points lower for the RNN, 0.91–5.41 percentage points lower for the Transformer, and 2.59–4.99 percentage points lower for the REINVENT prior.

Fixed-budget usable yield showed a generator-dependent response to NE (Figure 4d–f for the metal-binding objective). Applying NE to all RNN parameters produced usable yields 3.42–6.56 percentage points below positive fine-tuning, primarily because its valid-output fraction was 7.03–12.31 percentage points lower; uniqueness and novelty among valid outputs remained similar. By contrast, scoped Transformer NE increased usable yield by 6.90–35.65 percentage points relative to positive fine-tuning and retained 88.1–91.6% validity, close to the 93.2% base value.

For the REINVENT prior, target-hit fractions decreased from 5.0–10.2% at baseline to 1.5–2.6% under NE, and usable yield increased by 2.92–4.88 percentage points relative to positive fine-tuning. The resulting REINVENT usable yields were 95.3–96.9%, compared with 88.3–93.4% for the base prior. For every architecture–objective combination, the paired differences between NE and positive fine-tuning in both target-hit fraction and usable yield remained significant after Holm correction (Supplementary Table S3; Supplementary Figure S10).

To test whether positive fine-tuning was disadvantaged by the selected training duration, we repeated the reactive-motif experiment for up to ten epochs using development seed 11 (Supplementary Figure S6). Longer positive fine-tuning improved motif removal in all three generators, but even after ten epochs it did not match the motif-removal effect of standard NE in any of them. In the RNN, positive fine-tuning nevertheless produced a higher usable yield because the stronger motif removal under full-model NE was offset by reduced validity. This validity loss might be mitigated by restricting the edited parameters, as demonstrated for the Transformer, although this was not tested for the RNN. Thus, prolonged positive fine-tuning improved the target endpoint but did not consistently remove the trade-off between motif removal and structural diversity.

We analyzed the metal-binding-motif-removal objective in more detail as a representative structural-diversity case study. Across all three generators, NE produced a higher unique-scaffold fraction among valid generated molecules than positive fine-tuning (Figure 4g–i). Positive fine-tuning narrowed the RNN and Transformer outputs relative to their base generators, whereas NE maintained or increased both scaffold coverage and mean pairwise Tanimoto distance (Supplementary Figure S8). The REINVENT prior retained high scaffold diversity under both interventions, with the highest unique-scaffold fraction under NE.

Distributional-distance analysis showed a measurable redistribution of chemical space (Supplementary Figure S9). Relative to the unfiltered base distribution, the mean Fréchet ChemNet distance (FCD) ± standard deviation across seeds under NE was 6.14 ± 0.86 for the RNN, 8.66 ± 0.85 for the Transformer, and 1.97 ± 0.11 for the REINVENT prior. For the RNN and Transformer, using the motif-free subset of the corresponding base samples as the reference reduced these distances to 4.70 ± 0.74 and 6.59 ± 0.57, respectively, indicating that part of the displacement resulted directly from removing molecules containing the selected motifs. Because FCD and Fréchet descriptor distance (FDD) do not have universal thresholds for defining small or large distributional shifts, we also compared central and extreme molecular-weight subsets with unfiltered draws from each generator's saved base pool (Supplementary Figure S7). Equal query and reference counts were used across all molecular-weight bands within each seed. Extreme molecular-weight restrictions produced larger distances than central bands, but the absolute values, particularly FCD, depended on sample size. These controls therefore illustrate the effect of chemical-space restriction rather than providing numerical thresholds for the model comparisons, which used different sample counts. Together with the scaffold results, the analyses show that NE suppressed the targeted chemistry without the reduction in scaffold diversity observed under positive fine-tuning.

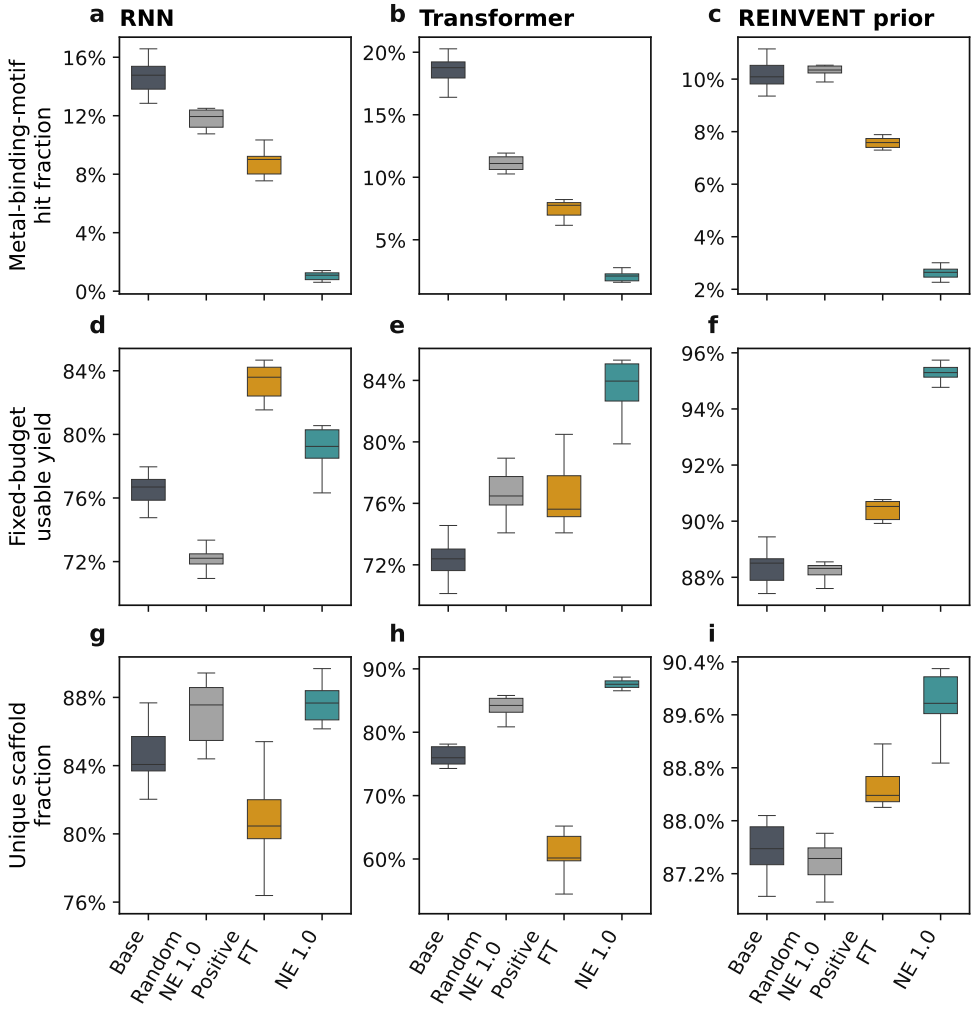


Fig. 4 Matched metal-binding-motif-removal experiment across three string-based molecular generators and ten paired confirmatory seeds, with 10,000 requested samples per model and seed. Columns show the RNN (a,d,g), the scoped Transformer (b,e,h), and the public pretrained REINVENT prior (c,f,i). The top row reports the metal-binding-motif hit fraction among valid outputs, the middle row the fixed-budget usable yield among all requested samples, and the bottom row the unique non-empty Bemis–Murcko scaffold count divided by the number of valid generated molecules. 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. Boxes summarize ten seed-level values; whiskers extend to 1.5 times the interquartile range. Results for the other three motif objectives and all lambda-response curves are reported in Supplementary Figures S3 and S4, respectively.

3.4 Joint motif suppression improves usable yield in a three-dimensional flow-matching generator

The preceding experiments established that NE can suppress targeted motifs across custom and pretrained string-based generators, although its effect on generative quality depended on the architecture. Molecular generators increasingly operate on molecular graphs and three-dimensional structures rather than string representations alone. We therefore investigated whether NE-based motif suppression transfers to SemlaFlow, an equivariant flow-matching model that jointly generates molecular graphs and three-dimensional conformations. We selected SemlaFlow because an independent comparison identified it as the strongest overall performer among five recent unconditional three-dimensional generators [11]. This experiment therefore extends the evaluation to a more complex output modality and a strong public baseline. Exploratory screening showed that restricting the SemlaFlow update to smaller parameter subsets preserved graph validity but substantially weakened liability removal (Supplementary Figure S5b). We therefore retained full-model extrapolation and investigated correction directions as an alternative way to preserve generative quality. We jointly targeted all four liability families. The base generator produced at least one target liability in 31.1% of valid molecules. Positive fine-tuning on liability-free molecules reduced this frequency to 13.3%, whereas standard NE reduced it to 5.3% (Figure 5a). The corresponding norm-matched random NE control reached only 25.7%, showing that an arbitrary trained direction of comparable magnitude did not reproduce the targeted response.

Standard full-model NE reduced RDKit-valid graph yield from 93.9% to 65.9% (Figure 5b), although the conditional PoseBusters pass rate among surviving valid graphs remained similar to the base generator (92.4% versus 91.7%; Figure 5c). To preserve liability removal while recovering graph validity, we formed corrected NE (CNE)

by subtracting a task vector learned from liability-free molecules from the negative-set task vector before extrapolation. The liability-free vector was used as a correction direction intended to remove shared changes associated with continued training on valid chemistry, rather than as a separate enrichment objective. CNE restored RDKit-valid graph yield to 88.3% while reducing liability frequency to 2.2% (Figure 5a,b). This suggests that subtracting the liability-free update canceled the part shared by both fine-tuning directions that supports valid generation, while retaining the difference associated with the targeted motifs. Paired effect estimates and exact paired-test results are reported in Supplementary Figure S12 and Supplementary Table S5.

For the unedited base generator, three-dimensional usable yield corresponds to the fraction retained after applying the same validity, uniqueness, liability, and PoseBusters criteria by post-generation filtering. Three-dimensional usable yield was 60.0% for the base generator, 76.6% after positive fine-tuning, 57.9% after standard NE, and 80.8% after CNE (Figure 5d). Relative to positive fine-tuning, CNE reduced liability frequency by 11.11 percentage points (95% CI, 10.82–11.40) and increased three-dimensional usable yield by 4.17 percentage points (95% CI, 3.42–4.93); both differences remained significant after Holm correction (Supplementary Table S5). Positive fine-tuning increased usable molecular yield but concentrated the resulting chemistry: the unique-scaffold fraction among usable molecules decreased from 88.6% for the base generator to 80.7%, whereas CNE reached 93.7% (Figure 5f). Consequently, CNE produced 3,786 unique three-dimensional usable scaffolds per 5,000 samples (75.7%), compared with 3,093 (61.9%) after positive fine-tuning (Figure 5g), and required fewer samples to obtain 2,500 usable scaffolds (3,279 versus 4,011; Figure 5h). Correction also reduced the mean FCD to the base distribution from 10.12 ± 0.40 for standard NE to 4.75 ± 0.53 for CNE (mean $\pm$ standard deviation across seeds; Figure 5e). Thus, CNE converted strong liability removal into improved molecular and scaffold yield while limiting the accompanying displacement of the generated distribution.

Overall, these results show that NE can substantially improve fixed-budget generative efficiency when architecture-specific validity losses are controlled. For the case study of SemlaFlow, this meant combining the standard NE update with a correction update to mitigate validity collapse.

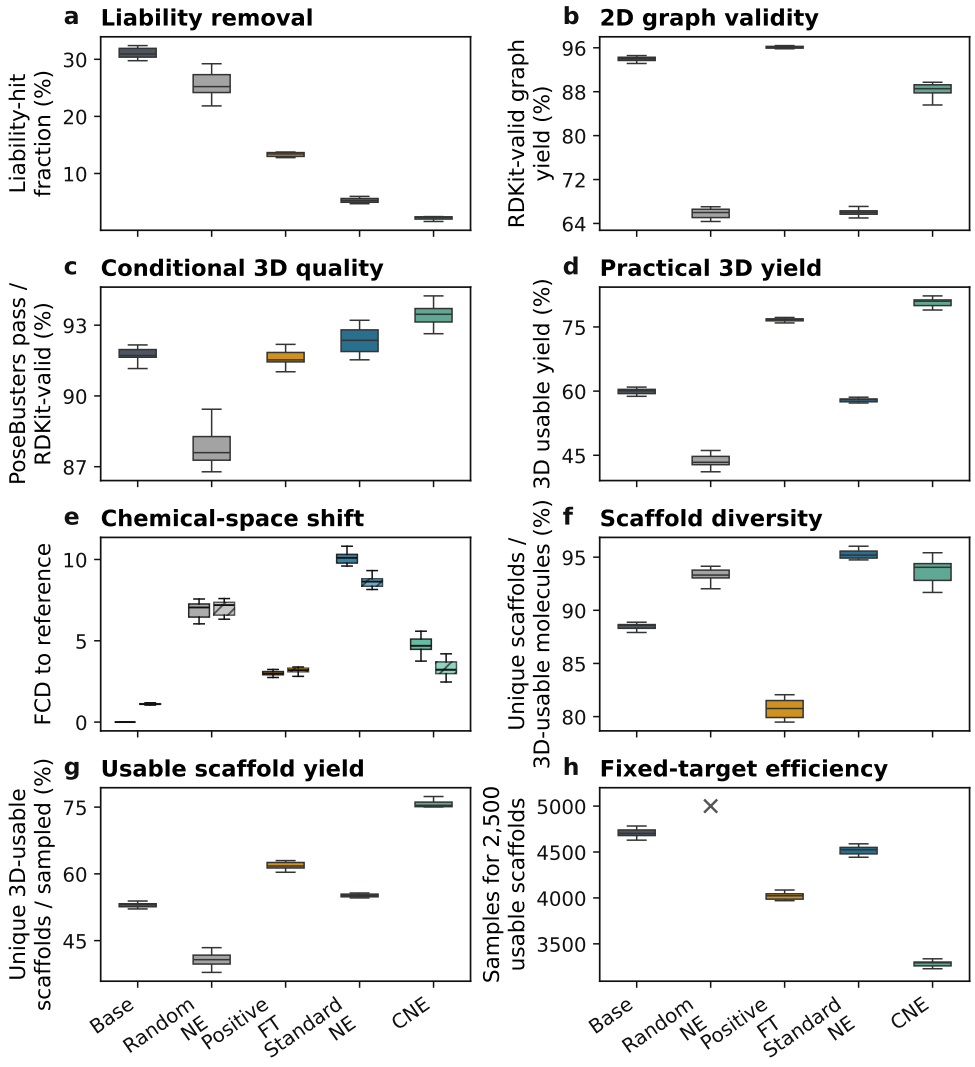


Fig. 5 Joint removal of all four motif classes in the pretrained SemlaFlow three-dimensional generator across ten paired confirmatory seeds. Panels show (a) joint target-motif hit fraction among RDKit-valid outputs, (b) RDKit-valid graph yield among all requested samples, (c) conditional PoseBusters pass rate among RDKit-valid outputs, and (d) three-dimensional usable yield among all requested samples. Three-dimensional usable molecules are RDKit-valid, canonically unique, free of all four target-motif classes, and pass every PoseBusters check. Panel (e) shows FCD to independently sampled raw-base (solid boxes) and motif-free-base (hatched boxes) references, (f) the unique Bemis–Murcko scaffold fraction among three-dimensional usable molecules, (g) unique three-dimensional usable scaffolds per requested sample, and (h) the number of samples required to obtain 2500 unique three-dimensional usable scaffolds. In (h), the × at 5,000 samples indicates that random NE did not reach the target within this budget in any of the ten seeds; it does not represent a measured target-completion count. Corrected NE (CNE) subtracts a task vector learned from liability-free valid molecules from the negative-set task vector before extrapolation. Boxes summarize ten seed-level values; whiskers extend to 1.5 times the interquartile range.

4 Discussion

The experiments conducted here support a specific role for NE in molecular generation. NE is not a universal optimizer: positive fine-tuning outperformed NE in the QED control, where high-scoring examples were readily available. Its value appears instead in settings where the undesirable region is easier to define than the desired region. Targeted structural liabilities satisfy this condition: metal-binding motifs, charged motifs, reactive groups, and assay-interference patterns are relatively sharp negative classes, whereas the corresponding "good" chemistry is broad and heterogeneous.

This interpretation is consistent across the RNN, Transformer, REINVENT prior, and SemlaFlow, but the fixed-budget benefit depended on architecture. NE reduced target-motif frequency more strongly than positive fine-tuning and norm-matched trained-random controls in all confirmatory liability experiments. This stronger suppression increased usable yield for the scoped Transformer, REINVENT prior, and corrected SemlaFlow models. Applying NE to all RNN parameters, however, produced lower usable yield than positive fine-tuning despite stronger liability removal, primarily because validity decreased rather than because valid outputs lost uniqueness or novelty. The later Transformer and SemlaFlow experiments show that update restriction or correction can recover architecture-specific validity losses. Applying analogous stabilization to the RNN may therefore improve its fixed-budget yield, but this was not tested. Structural-diversity analyses further suggest that NE does not simply concentrate the generator on a narrow favorable subset. Instead, it can move the generator away from an undesirable region while maintaining a broad output distribution. This distinction is important for medicinal chemistry, where avoiding a liability is often not equivalent to converging on one positive chemotype.

The Transformer and SemlaFlow experiments revealed distinct architecture-dependent failure modes. Full-model extrapolation severely reduced Transformer SMILES validity, whereas restricting the update to the final Transformer block and

output layer preserved validity during liability removal. In SemlaFlow, full-model NE retained three-dimensional quality among RDKit-valid molecules but reduced the overall yield of valid molecular graphs. The appropriate intervention was therefore also architecture-dependent: scope restriction stabilized the Transformer, whereas task-vector correction stabilized SemlaFlow.

The SemlaFlow correction experiment provides a possible explanation for the loss of validity under standard NE. Fine-tuning on liability-containing molecules may produce a task vector that contains both information associated with the targeted liabilities and more general changes that support generation of valid molecules. Reversing the complete vector can therefore suppress the liabilities while also degrading validity. A related problem is described as incomplete weight disentanglement in task arithmetic, where a parameter update also affects model behavior outside its intended target [25]. Using the liability-free task vector as a correction direction recovered most of the lost graph validity without weakening liability removal. This is consistent with validity-related changes shared by both task vectors cancelling when the liability-free update was subtracted from the negative-set update. This interpretation is supported by the model outputs but does not establish a unique decomposition of the parameter updates; direct analyses of vector alignment and shared parameter subspaces would be needed to test the proposed mechanism at the weight level.

Positive and negative parameter directions have previously been combined for few-shot molecular optimization [29]. In our SemlaFlow experiment, the liability-free direction instead served as a correction for the validity loss caused by reversing the complete negative-set update. The contrast between standard and corrected NE also shows why target-hit frequency alone is insufficient for evaluating model edits; validity, chemical diversity, and usable molecular and scaffold yield must be considered alongside the targeted response.

The reduction in motif frequency should not be interpreted as concept erasure or machine unlearning. Our experiments measure changes in the generated distribution, not whether the model retains the underlying ability to produce a motif. We did not test whether subsequent fine-tuning or other interventions could restore motif generation.

Several limitations remain. First, the liability definitions used here are SMARTS-based structural proxies, not experimentally validated toxicity or developability measurements. Their utility is therefore to test controllable motif suppression, not to establish that every matched molecule would fail experimentally. Second, we evaluated NE in four generator implementations spanning recurrent, Transformer, and three-dimensional flow-matching architectures, but this still represents only a small part of the molecular-generation landscape. Models trained on other molecular representations and datasets, or edited for less clearly defined objectives, may respond differently. Broader testing will therefore be needed to establish which model classes and applications benefit from NE, and where the method is ineffective. Third, the present work studies unconditional generation. Conditional ligand- or structure-based design may couple the liability direction to conditioning information and require different update scopes or training losses. Finally, distributional metrics such as FCD and FDD are useful for auditing chemical-space movement, but do not by themselves define whether a shift is medicinally impactful.

5 Conclusion

NE provides a simple parameter-space strategy for selectively suppressing undesirable chemistry by learning a direction toward liability-containing molecules and extrapolating away from it. Across custom recurrent and Transformer generators, a public pretrained recurrent prior, and a three-dimensional flow-matching generator, NE consistently reduced targeted structural liabilities. Beyond demonstrating

this effect, our experiments identify two ways to adapt parameter-space editing to different model architectures. Restricting the edit to late Transformer parameters preserved molecular validity while retaining liability removal, whereas CNE combined negative and motif-free update directions to preserve liability removal while restoring valid graph generation in SemlaFlow. In the three-dimensional generator, this correction also increased the fixed-budget yield of usable scaffolds. Across the representative structural-diversity analyses, NE-based editing produced higher usable-scaffold diversity than positive fine-tuning, showing that stronger motif removal did not require a reduction in chemical diversity. These findings establish parameter scope and correction-direction construction as practical extensions of NE, rather than treating architecture-dependent validity loss as an unavoidable limitation. Together, the results define NE not as a general-purpose molecular optimizer, but as a targeted model-editing approach for objectives that are more naturally specified by what should be excluded than by what should be enriched. Future work should test this principle in conditional and structure-guided generators and with experimentally validated medicinal-chemistry objectives.

6 Methods

6.1 RNN and Transformer generators

We implemented the RNN and Transformer as two unconditional autoregressive SMILES generators in PyTorch [36] and trained both on the GuacaMol training set [31]. This standard distribution-learning split contains 1,273,104 molecular SMILES derived from ChEMBL 24 [32]. The first generator was an RNN, comprising a token-embedding layer, a three-layer gated recurrent unit network, and a linear vocabulary projection [37]. The second generator was a Transformer with learned token and positional embeddings, four causally masked Transformer blocks, layer normalization, and a linear vocabulary projection [38]. Both models used the same SMILES tokenizer and

were trained to predict the next token using padding-masked categorical cross-entropy and AdamW [39]. Detailed dimensions and training hyperparameters are provided in Supplementary Section 1.2.

Both base architectures were trained for eight epochs on the full GuacaMol training set. This fixed training budget was selected during base-model development, before any downstream objective was evaluated, and was held constant across all seeds. Base training was independently repeated for confirmatory seeds 13, 17, 19, 23, 29, 31, 37, 41, 43, and 47. Exploratory method development used seed 11; seeds 5 and 7 were additionally used only for the Transformer scope ablation. Results from seeds 5, 7, and 11 were excluded from confirmatory summaries, hypothesis tests, and confirmatory figures.

Each objective-specific experiment started from the corresponding saved base checkpoint for that seed (Supplementary Section 1.3). Each checkpoint generated 30,000 initial molecules for objective-set selection and 10,000 molecules for endpoint evaluation. Objective-specific fine-tuning used two epochs for the RNN and one epoch for the Transformer. These short updates were chosen to estimate local parameter directions while limiting drift from the base generator; an exploratory 1–10-epoch sensitivity analysis is reported in Supplementary Figure S6. The RNN liability experiments used full-model NE. The final Transformer liability experiments modified only the fourth and final Transformer block together with the output projection that maps hidden representations to token probabilities (parameter prefixes `transformer.layers.3.` and `output.`). All embedding layers and earlier Transformer blocks remained at their base-model values during extrapolation. This scope was selected in an exploratory ablation on metal-binding-motif removal using development seeds 5, 7, and 11, with every parameter scope evaluated at $\lambda \in \{0.1, 0.25, 0.5, 0.75, 1.0\}$ (Supplementary Section 1.5).

6.2 Objectives and liability definitions

We evaluated two classes of objective. First, we used QED as a direct molecular analogue of the original NEON setting: molecules in the lower tail of the QED distribution represented lower-quality generated data, and reversing the update learned from them was intended to improve overall output quality. High-QED molecules also provide a clearly defined positive fine-tuning set, making this a useful comparison between NE and conventional fine-tuning. We then considered objectives that are more naturally defined by negative examples. A molecule can be assigned to an unwanted class through the presence of a specific structural motif, whereas the molecules without that motif may otherwise be chemically unrelated. The four binary motif-removal experiments therefore tested whether NE could selectively suppress a well-defined negative region without concentrating the generator toward a narrow positive subset.

For the QED objective [35], the bottom 10% of valid base-generated molecules formed the negative set and the top 10% formed the positive set. Motif-removal objectives used binary structural flags computed with RDKit [40]. The main string-based-generator analyses used four objective-specific flags: reactive motifs, metal-binding motifs, charged motifs, and assay-interference motifs [12–16]. For each objective, molecules matching the target flag formed the negative set, and molecules not matching the target flag formed the positive set. Selected sets were balanced 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 SemlaFlow experiment used a joint flag defined as the logical union of the same four motif classes. Its negative set was balanced across classes to prevent the most frequent class from dominating the update. The objective rationale and score columns are summarized in Supplementary Section 1.4 and Supplementary Table S1; exact SMARTS patterns are given in Supplementary Table S2 and representative matches in Supplementary Figure S1.

For each generated set, validity was defined as successful RDKit parsing, sanitization, and canonicalization. Uniqueness was the fraction of unique canonical SMILES among valid outputs. For the RNN and Transformer, novelty was assessed against the canonicalized GuacaMol training set [31]; no external novelty reference was assumed for public pretrained models whose complete training molecules were unavailable. QED was computed with RDKit's `QED.qed` [35, 40]; molecular weight, Crippen logP [41], topological polar surface area (TPSA) [42], hydrogen-bond donors and acceptors, formal charge, ring counts, aromatic ring counts, and fraction of sp^3 carbons were computed with RDKit descriptor functions. Motif-hit fractions were computed from objective-specific SMARTS flags. Scaffold diversity summaries were computed with RDKit using Bemis–Murcko scaffolds [43] and Morgan fingerprints [44] with radius 2 and 2048 bits [40]. Inspired by fixed-budget valid–unique–novel metrics used to evaluate molecular generators [11, 45], we defined an extended usable-yield endpoint:

$$\text{usable yield} = \frac{N_{\text{acceptable}}}{N_{\text{requested}}}, \tag{6}$$

where the acceptance criteria depended on the available reference data and the generated representation (Supplementary Sections 1.11 and 1.10). For the RNN and Transformer motif-removal tasks, acceptable molecules were valid, canonically unique, novel relative to the GuacaMol training set, and free of the target motif. For the public REINVENT prior, novelty was omitted because a complete canonical training-set reference was unavailable. This metric reports the fraction of the fixed sampling budget meeting the generator-specific acceptance criteria.

6.3 REINVENT prior experiments

To test transfer to a public pretrained model, we used the classical recurrent REINVENT prior distributed with REINVENT4 at

`external/REINVENT4/priors/reinvent.prior` [20, 46]. The ChEMBL-trained checkpoint has a 256-dimensional embedding and three long short-term memory (LSTM) layers of hidden size 512. For each seed and motif-removal objective, we sampled 50,000 baseline molecules for training-set selection and selected balanced negative and positive sets as above. Each edited model and the unedited prior were evaluated with a separate budget of 10,000 requested samples per seed. The same seed-specific unedited-prior evaluation was scored for all four objectives; the 50,000-molecule pools remained the distributional-distance and molecular-weight calibration references. REINVENT transfer learning used two epochs, batch size 256, learning rate 10^{-4}, randomized SMILES, and unique-molecule sampling. Because the purpose of transfer learning was to estimate a local parameter-update direction rather than fully specialize the prior, negative-set, motif-free, and random-set updates were each limited to two epochs. This duration was fixed before the confirmatory experiments; sensitivity to updates of up to ten epochs is reported in Supplementary Figure S6. The NE lambda grid was 0.25, 0.5, 0.75, and 1.0. We applied full-model NE to the REINVENT prior rather than a scoped update. Molecules containing tokens unsupported by the prior vocabulary were excluded from transfer-learning input files. Further implementation details are provided in Supplementary Section 1.6.

6.4 SemlaFlow three-dimensional generator

To evaluate transfer to three-dimensional generation, we used the public SemlaFlow checkpoint trained for 200 epochs on GEOM-Drugs [30, 33]. SemlaFlow was selected because it showed the strongest overall performance in an independent comparison of five recent unconditional three-dimensional generators and because its implementation and pretrained weights were publicly available [11]. SemlaFlow jointly generates atom coordinates, atom types, bond types, and formal charges using equivariant flow matching. The checkpoint uses a 384-dimensional model, 12 equivariant layers, 32 attention

heads, 64 coordinate sets, and a maximum of 256 atoms. Sampling used the published exponential-moving-average parameters, 100 integration steps, logarithmic ordinary differential equation (ODE) sampling, and categorical noise level one (Supplementary Section 1.7).

For each confirmatory seed (13, 17, 19, 23, 29, 31, 37, 41, 43, and 47), SemlaFlow generated an independent 50,000-molecule base pool. The joint negative set contained 3000 training molecules and 300 validation molecules assigned to each of the four motif classes, giving 12,000 training and 1200 validation examples while allowing molecules with multiple matches to be assigned only once. Matched random and motif-free sets used the same total sizes. Fine-tuning retained the SemlaFlow flow-matching objective and updated the full model for ten epochs at learning rate 3×10^{-4} (Supplementary Section 1.8). The ten-epoch duration, learning rate, full-model scope, and final $\lambda = 2.5$ were selected during exploratory screening on seed 11 with a separate oxygen–nitrogen single-bond proxy objective because shorter or lower-magnitude updates produced little target response. Seed 11 was excluded from the confirmatory analysis; the complete parameter-selection sequence is reported in Supplementary Figure S5 and Supplementary Section 1.9. Each model generated 5000 evaluation molecules per seed.

For SemlaFlow, RDKit-valid graph yield was the fraction of all sampled outputs that could be sanitized and processed as molecular graphs. Conditional three-dimensional quality was the fraction of RDKit-valid outputs passing every PoseBusters check under the `mol` configuration [34]. Three-dimensional usable yield was

$$\text{3D usable yield} = \frac{N_{\text{valid,unique,liability}-\text{free,PoseBusters}-\text{pass}}}{N_{\text{sampled}}}. \tag{7}$$

We additionally measured the fraction of unique Bemis–Murcko scaffolds among three-dimensional usable molecules, the number of unique three-dimensional usable scaffolds per sampling budget, and the number of samples required to obtain 2500 such scaffolds.

6.5 Structural diversity and distributional metrics

Structural diversity was summarized using the unique Bemis–Murcko scaffold fraction [43] and mean pairwise Morgan-fingerprint Tanimoto distance [44], where higher pairwise distance indicates lower average fingerprint similarity within a generated set. Distributional shift was quantified with FCD, comparing ChemNet embedding distributions [47], and FDD, introduced by Özçelik and Grisoni as a descriptor-space analogue of FCD [48]. Following their released implementation, the FDD descriptor vector consisted of logP, molecular weight, hydrogen-bond donors, ring count, and TPSA, scaled to fixed descriptor ranges before computing the Fréchet distance. Molecular-weight-biased subsets of saved base pools provided chemically interpretable examples of distribution shifts. We compared equal query and reference counts across all bands within each seed and examined sensitivity to alternative sample-count choices. Because the draws could overlap and the calibration counts differed from those of the model comparisons, these controls provide qualitative context rather than thresholds for small or large shifts (Supplementary Section 1.12).

6.6 Statistical analysis

Endpoint statistics were computed from ten paired confirmatory seeds, with each seed defining one experimental replicate. Generated molecules were not treated as independent replicates. For each comparison, we report the mean paired difference and its two-sided 95% Student-t CI. Hypothesis tests used an exact two-sided paired sign-flip test of the mean difference. Holm correction was used to account for multiple comparisons [49]. For each string-based generator and motif-removal endpoint, correction was applied jointly to the 12 tests arising from four motif-removal objectives and three

comparisons of NE: against base generation, positive fine-tuning, and norm-matched random NE. For the QED analysis, correction was applied separately to each architecture and endpoint across the same three comparisons. For SemlaFlow, correction was applied separately to each endpoint across five prespecified comparisons: base generation, positive fine-tuning, standard NE, random NE, and the norm-matched random-correction control (Supplementary Section 2.2). The complete testing procedure is described in Supplementary Section 1.13, with the paired results summarized in Supplementary Section 2.1. Effects for fractional endpoints are reported as percentage-point differences. Development-seed parameter screens, lambda-response curves, scope ablations, epoch-sensitivity experiments, and distributional analyses were treated as exploratory and were not included in confirmatory hypothesis tests.

Supplementary information

Additional file 1: Supplementary Information (PDF). Extended methods (Supplementary Section 1), supplementary results and statistical summaries (Supplementary Section 2), additional figures (Supplementary Section 3), and reproduction commands (Supplementary Section 4).

Declarations

Availability of data and materials

The GuacaMol and GEOM-Drugs datasets, the pretrained REINVENT prior distributed with REINVENT4, and the SemlaFlow checkpoint are publicly available from the original sources cited in the Methods. The generated molecule-level outputs, seed-level summary tables, statistical outputs, and figure-source data supporting this study are available in Zenodo [50] at https://doi.org/10.5281/zenodo.23165725. Source code,

experiment configurations, executable Makefile targets, analysis scripts, and figure-generation notebooks are available at https://github.com/JoshuaHesse/negative_extrapolation_molgen. The reproducibility workflow and computational resources are described in Supplementary Sections 1.1 and 1.14.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors thank Merck KGaA Darmstadt for their generous support with the Merck Future Insight Prize 2020. This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the KIbiotika project (03LWH0179).

Authors' contributions

Conceptualization: J.H.; Data curation: J.H.; Formal analysis: J.H.; Investigation: J.H.; Methodology: J.H.; Software: J.H.; Supervision: J.W. and S.A.S.; Writing original draft: J.H.; All authors have given approval to the final version of the manuscript.

Use of artificial intelligence tools

OpenAI Codex was used to assist with code implementation. ChatGPT (OpenAI) was used to refine sentence structure, grammar, and phrasing. These tools were not used for study conceptualization, hypothesis generation, experimental design, or scientific interpretation. All code and text produced with their assistance was reviewed and verified by the authors.

Acknowledgements

J.H. thanks Davide Boldini and Aleksandra Daniluk for their valuable feedback and discussions on this work.

References

[1] Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., Zentgraf, M., Hill, J.E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebkemann, C., Schneider, G.: Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery **19**(5), 353–364 (2020) https://doi.org/10.1038/s41573-019-0050-3 . Accessed 2026-08-10

[2] Elton, D.C., Boukouvalas, Z., Fuge, M.D., Chung, P.W.: Deep learning for molecular design—a review of the state of the art. Molecular Systems Design & Engineering **4**(4), 828–849 (2019) https://doi.org/10.1039/c9me00039a . Accessed 2026-08-10

[3] Meyers, J., Fabian, B., Brown, N.: *De novo* molecular design and generative models. Drug Discovery Today **26**(11), 2707–2715 (2021) https://doi.org/10.1016/j.drudis.2021.05.019 . Accessed 2026-08-10

[4] Walters, W.P., Murcko, M.: Assessing the impact of generative AI on medicinal chemistry. Nature Biotechnology **38**(2), 143–145 (2020) https://doi.org/10.1038/s41587-020-0418-2 . Accessed 2026-08-10

[5] Thomas, M., Astudillo, R., Raouf, Y.S., Navarro, C., De Fabritiis, G., Bender, A.: Generative Virtual Screening: From Search to Generate and Back (2026) https://doi.org/10.26434/chemrxiv.10001862/v3 . Accessed 2026-06-15

[6] Xu, H.-X., Peng, J., Yu, J.-L., Hong, X., Xiao, Y., Ye, B.-G., Li, G.-B.: Evolution of Molecular Generation Models in Drug Discovery. Journal of Medicinal Chemistry **69**(2), 845–866 (2026) https://doi.org/10.1021/acs.jmedchem.5c03310 . Accessed 2026-06-15

[7] Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science **4**(2), 268–276 (2018) https://doi.org/10.1021/acscentsci.7b00572 . Accessed 2026-08-10

[8] Segler, M.H.S., Kogej, T., Tyrchan, C., Waller, M.P.: Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Central Science **4**(1), 120–131 (2017) https://doi.org/10.1021/acscentsci.7b00512 . Accessed 2026-08-10

[9] Popova, M., Isayev, O., Tropsha, A.: Deep reinforcement learning for de novo drug design. Science Advances **4**(7), 7885 (2018) https://doi.org/10.1126/sciadv.aap7885 . Accessed 2026-08-10

[10] Ståhl, N., Falkman, G., Karlsson, A., Mathiason, G., Boström, J.: Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. Journal of Chemical Information and Modeling **59**(7), 3166–3176 (2019) https://doi.org/10.1021/acs.jcim.9b00325 . Accessed 2026-08-10

[11] Buttenschoen, M., Ziv, Y., Morris, G.M., Deane, C.M.: An evaluation of unconditional 3D molecular generation methods. arXiv. arXiv:2505.00518 [physics.chem-ph] (2025). https://doi.org/10.48550/arXiv.2505.00518 . http://arxiv.org/abs/2505.00518 Accessed 2026-06-15

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] Bruns, R.F., Watson, I.A.: Rules for Identifying Potentially Reactive or Promiscuous Compounds. Journal of Medicinal Chemistry **55**(22), 9763–9772 (2012) https://doi.org/10.1021/jm301008n . Accessed 2026-08-10

[19] Schuh, M.G., Hesse, J., Sieber, S.A.: AI-guided Antibiotic Discovery Pipeline from Target Selection to Compound Identification. arXiv. arXiv:2504.11091 [q-bio.BM] (2025). https://doi.org/10.48550/arXiv.2504.11091 . http://arxiv.org/abs/2504.11091 Accessed 2026-06-15

[20] Loeffler, H.H., He, J., Tibo, A., Janet, J.P., Voronov, A., Mervin, L.H., Engkvist, O.: Reinvent 4: Modern AI–driven generative molecule design. Journal of Cheminformatics **16**(1), 20 (2024) https://doi.org/10.1186/s13321-024-00812-5 . Accessed 2026-06-15

[21] Thomas, M., O'Boyle, N.M., Bender, A., De Graaf, C.: MolScore: a scoring, evaluation and benchmarking framework for generative models in de novo drug design. Journal of Cheminformatics **16**(1), 64 (2024) https://doi.org/10.1186/s13321-024-00861-w . Accessed 2026-08-10

[22] Westermayr, J., Gilkes, J., Barrett, R., Maurer, R.J.: High-throughput property-driven generative design of functional organic molecules. Nature Computational Science **3**(2), 139–148 (2023) https://doi.org/10.1038/s43588-022-00391-1 . Accessed 2026-09-16

[23] Gao, W., Fu, T., Sun, J., Coley, C.W.: Sample Efficiency Matters: A Benchmark for Practical Molecular Optimization. arXiv. arXiv:2206.12411 [cs.CE] (2022). https://doi.org/10.48550/arXiv.2206.12411 . http://arxiv.org/abs/2206.12411 Accessed 2026-08-10

[24] Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L.,

Hajishirzi, H., Farhadi, A.: Editing Models with Task Arithmetic. arXiv. arXiv:2212.04089 [cs.LG] (2023). https://doi.org/10.48550/arXiv.2212.04089 . http://arxiv.org/abs/2212.04089 Accessed 2026-08-10

[25] Ortiz-Jimenez, G., Favero, A., Frossard, P.: Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. arXiv. arXiv:2305.12827 [cs.LG] (2023). https://doi.org/10.48550/arXiv.2305.12827 . http://arxiv.org/abs/2305.12827 Accessed 2026-08-10

[26] Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing Concepts from Diffusion Models. arXiv. arXiv:2303.07345 [cs.CV] (2023). https://doi.org/10.48550/arXiv.2303.07345 . http://arxiv.org/abs/2303.07345 Accessed 2026-08-10

[27] Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., Zhu, J.-Y.: Ablating Concepts in Text-to-Image Diffusion Models. arXiv. arXiv:2303.13516 [cs.CV] (2023). https://doi.org/10.48550/arXiv.2303.13516 . http://arxiv.org/abs/2303.13516 Accessed 2026-08-10

[28] Alemohammad, S., Wang, Z., Baraniuk, R.G.: Neon: Negative Extrapolation From Self-Training Improves Image Generation. arXiv. arXiv:2510.03597 [cs.GR] (2025). https://doi.org/10.48550/arXiv.2510.03597 . http://arxiv.org/abs/2510.03597 Accessed 2026-06-15

[29] Özçelik, R., Ruiter, S.d., Grisoni, F.: Look the Other Way: Designing 'Positive' Molecules with Negative Data via Task Arithmetic (2025). Accessed 2026-06-15

[30] 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

[31] 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

[32] Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M., Mosquera, J., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C., Segura-Cabrera, A., Hersey, A., Leach, A.: ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research **47**(D1), 930–940 (2019) https://doi.org/10.1093/nar/gky1075 . Accessed 2026-09-16

[33] Axelrod, S., Gomez-Bombarelli, R.: GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. arXiv. arXiv:2006.05531 [physics.comp-ph] (2022). https://doi.org/10.48550/arXiv.2006.05531 . http://arxiv.org/abs/2006.05531 Accessed 2026-08-10

[34] Buttenschoen, M., Morris, G.M., Deane, C.M.: PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chemical Science **15**(9), 3130–3139 (2024) https://doi.org/10.1039/d3sc04185a . Accessed 2026-08-10

[35] Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., Hopkins, A.L.: Quantifying the chemical beauty of drugs. Nature Chemistry **4**(2), 90–98 (2012) https://doi.org/10.1038/nchem.1243 . Accessed 2026-08-10

[36] 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

[37] 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

[38] 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

[39] 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

[40] Landrum, G., Tosco, P., Kelley, B., Ric, Cosgrove, D., sriniker, Vianello, R., gedeck, NadineSchneider, Jones, G., Kawashima, E., N, D., Dalke, A., Cole, B., Swain, M., Turk, S., Savelev, A., Vaucher, A., Wójcikowski, M., Take, I., Scalfani, V.F., Probst, D., Ujihara, K., godin, g., Pahl, A., Walker, R., Lehtivarjo, J., Berenger, F., strets123, jasondbiggs: rdkit/rdkit: Release_2023.09.5. Zenodo (2024). https://doi.org/10.5281/zenodo.10633624 . https://doi.org/10.5281/zenodo.10633624

[41] Wildman, S.A., Crippen, G.M.: Prediction of Physicochemical Parameters by Atomic Contributions. Journal of Chemical Information and Computer Sciences **39**(5), 868–873 (1999) https://doi.org/10.1021/ci9903071 . Accessed 2026-08-10

[42] Ertl, P., Rohde, B., Selzer, P.: Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. Journal of Medicinal Chemistry **43**(20), 3714–3717 (2000) https://doi.org/10.1021/jm000942e . Accessed 2026-08-10

[43] Bemis, G.W., Murcko, M.A.: The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry **39**(15), 2887–2893 (1996) https://doi.org/10.1021/jm9602928 . Accessed 2026-08-10

[44] Rogers, D., Hahn, M.: Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling **50**(5), 742–754 (2010) https://doi.org/10.1021/ci100050t . Accessed 2026-08-10

[45] Boget, Y., Gregorova, M., Kalousis, A.: Graph annotation generative adversarial networks. In: Khan, E., Gonen, M. (eds.) Proceedings of the 14th Asian Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 189, pp. 16–16. PMLR, Hyderabad, India (2023). https://proceedings.mlr.press/v189/boget23a.html

[46] Olivecrona, M., Blaschke, T., Engkvist, O., Chen, H.: Molecular de-novo design through deep reinforcement learning. Journal of Cheminformatics **9**(1), 48 (2017) https://doi.org/10.1186/s13321-017-0235-x . Accessed 2026-06-15

[47] Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S., Klambauer, G.: Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery. Journal of Chemical Information and Modeling **58**(9), 1736–1741 (2018)

https://doi.org/10.1021/acs.jcim.8b00234 . Accessed 2026-08-10

[48] 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

[49] Holm, S.: A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics **6**(2), 65–70 (1979). Accessed 2026-08-10

[50] Hesse, J.: Negative Extrapolation for Molecular Generation: Generated Molecules and Analysis Data. Zenodo. Dataset (2026). https://doi.org/10.5281/zenodo.23165725 . https://doi.org/10.5281/zenodo.23165725

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NegativeextrapolationformoleculargeneratorsSIsubmission.pdf