How Creative Are Large Language Models in Generating Molecules?

Anonymous authors
Paper under double-blind review

Abstract

Molecule generation requires satisfying chemical and biological constraints while searching a large and structured chemical space. This makes molecule generation a non-binary problem: many distinct molecular structures can satisfy the same constraints, and effective models must identify non-obvious solutions while maintaining exploration. From this perspective, creativity is a functional requirement in molecular generation rather than an aesthetic notion. Large language models (LLMs) can generate molecular representations directly from natural language prompts, but it remains unclear what kind of creativity they exhibit when generating molecules. Through a systematic evaluation of 10 general-purpose LLMs across 15 molecular generation tasks, we study creativity along two complementary dimensions: convergent creativity, which captures validity and constraint satisfaction, and divergent creativity, which captures novelty, uniqueness, and structural exploration. Our central finding is that higher convergent creativity is consistently associated with lower divergent creativity: the relationship is negative within all 15 evaluation tasks. We next examine how molecular creativity varies with three factors: task structure, model family, and inference-time conditioning. In the tested compatible multi-constraint tasks, adding constraints does not necessarily reduce success; across size-varied checkpoints, the LLaMA families gain convergent creativity while losing divergent creativity, whereas Qwen3 improves both; and in-context demonstrations can sharply increase activity-task success while reducing exploration and directly reproducing supplied examples. We further test whether creativity enhanced in language transfers to molecular generation. In a controlled same-parent comparison, the three creativity-oriented variants fail to reproduce their targeted molecular gains.

1 Introduction

Designing new functional molecules aims to identify chemical structures that satisfy desired properties, such as selective inhibition of a disease target, while also meeting additional constraints related to stability, synthesizability, and safety (Gao et al., 2022). For a given request, there often exist many chemically valid solutions. Molecule generation is therefore not a binary problem with strictly correct or incorrect answers, and what distinguishes strong models is whether they can discover non-obvious structures that satisfy the constraints, rather than reproducing close variants of known examples. Given the high failure rate in drug discovery (DiMasi et al., 2010), generating diverse designs is important for mitigating failure risk and navigating intellectual property restrictions.

Molecule generation therefore requires two complementary abilities: satisfying chemical and functional constraints, and exploring diverse regions of chemical space for alternative viable solutions. We refer to this combination as creativity in a functional sense. Under this view, a creative molecular generator should not only produce valid molecules that satisfy the requested properties, but also avoid collapsing onto a narrow set of familiar structures. Here, creativity is treated as a functional property of molecular generation rather than a subjective or aesthetic notion. The challenge is therefore not simply to maximize constraint satisfaction or diversity in isolation, but to understand how the two behave together.

This question becomes particularly relevant as molecular generation shifts from specialized generative models to general-purpose LLMs. Molecular generation has traditionally been approached with

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

models designed specifically for chemical structures, such as variational autoencoders (Jin et al., 2018) and diffusion-based models (Hoogeboom et al., 2022). LLMs introduce a different paradigm: they can generate molecular representations directly from natural-language prompts, combine multiple constraints without retraining, and adapt their outputs through prompting or in-context examples (Li et al., 2024c; Wang et al., 2023; Tao et al., 2025). This flexibility makes LLMs particularly appealing for molecular design, but it remains unclear how constraint satisfaction and chemical-space exploration change together as these models are prompted, scaled, or otherwise adapted.

We study this question through two complementary dimensions of molecular creativity: convergent creativity (CC), which captures validity and constraint satisfaction, and divergent creativity (DC), which captures novelty, uniqueness, and structural exploration. We conduct a systematic empirical study of 10 general-purpose LLMs across 15 molecular generation tasks spanning physicochemical, ADMET, biological activity, and multi-constraint objectives.

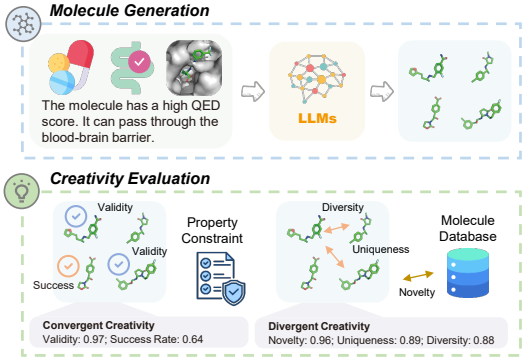


Figure 1: Overview of how LLM creativity in molecular generation is evaluated. **Top:** LLMs generate molecules conditioned on physicochemical, ADMET, and biological activity constraints specified in the prompt. **Bottom:** Creativity is operationalized along two complementary dimensions: convergent creativity (constraint satisfaction) and divergent creativity (chemical space exploration).

Our central finding is a recurring negative association between convergent and divergent creativity: the relationship is negative within all 15 evaluation tasks and remains after accounting for task-family and model-specific baselines. This means that evaluating only constraint satisfaction or only exploration can miss systematic changes in the other dimension.

We then examine what shapes these two dimensions. Task families occupy distinct regions of the CC–DC plane; across size-varied checkpoints, the LLaMA families gain CC while losing DC, whereas Qwen3 improves both; and ten-shot in-context learning sharply increases activity-task success while reducing exploration and substantially reusing the supplied examples. Finally, we ask whether creativity-oriented preference tuning in language transfers to molecular generation. In a controlled same-parent comparison, none of the three creativity-oriented variants produces the molecular gain targeted by its language-domain objective. Table 1 summarizes the main findings, practical guidance, and supporting evidence.

2 Related Work

Molecular generation and evaluation. Molecular generation has been extensively studied with specialized generative models, including variational autoencoders (Gómez-Bombarelli et al., 2018; Simonovsky & Komodakis, 2018), generative adversarial networks (Guimaraes et al., 2017; De Cao & Kipf, 2018), flow-based models (Shi et al., 2020; Song et al., 2023), autoregressive and 3D generative models (Popova et al., 2019; Gebauer et al., 2019), and diffusion models (Hoogeboom et al., 2022; Xu et al., 2022). Prior molecular-generation work commonly optimizes molecular properties (Jin et al., 2020b; Gao et al., 2022; Lee et al., 2025) and evaluates validity, property satisfaction, novelty, uniqueness, and structural diversity (Polykovskiy et al., 2020). However, these metrics are typically evaluated independently, leaving the joint behavior of constraint satisfaction and molecular exploration largely uncharacterized.

LLMs for molecular generation. Recent work has shown that general-purpose LLMs can generate molecular structures from natural-language instructions (Li et al., 2024c; Tao et al., 2025), molecular prompts (Wang et al., 2023), and in-context examples (Tao et al., 2025). Existing studies mainly ask whether LLMs can generate valid molecules and satisfy target properties (Li et al., 2024b; Tao et al., 2025). We instead study how constraint satisfaction and molecular exploration change together, and how this relationship varies across molecular tasks, model families, and inference-time conditions.

Table 1: Main findings and practical implications. Molecular creativity should be evaluated jointly along constraint satisfaction and exploration, as both vary with the task, model, and generation strategy.

Theme	Main finding	Practical takeaway	Where to look
Core relationship	CC and DC show a recurring negative association: all 15 within-task correlations are negative, and the relationship remains after accounting for task-family and model differences.	**Evaluate constraint satisfaction and exploration together.** Looking at either dimension alone can hide systematic changes in the other.	Sec. 4; Fig. 2
Task structure	Task families occupy distinct regions of the CC–DC plane. In the tested compatible multi-constraint tasks, adding a second constraint does not necessarily reduce success.	**Expect different CC–DC profiles across tasks.** Performance and exploration observed on one task family need not carry over to another.	Sec. 5; Fig. 3A
Model family	Across size-varied checkpoints, the LLaMA families gain CC while losing DC, whereas Qwen3 improves both dimensions.	**Do not infer molecular creativity from model size alone.** Compare both CC and DC on the target task.	Sec. 5.2; Fig. 3B
Inference-time control	Ten-shot ICL sharply increases activity-task success while reducing DC and substantially reusing supplied examples. Temperature and diversity prompting also shift CC and DC in task- and model-dependent ways.	**Inference-time controls can reshape the CC–DC profile without retraining.** Monitor exploration and copying alongside gains in constraint satisfaction.	Sec. 5.3; Fig. 3C
Cross-domain transfer	Creativity-oriented preference tuning in language does not reproduce its targeted gains in molecular generation. CC decreases across all four preference-tuned variants, with reduced validity accounting for most of the decline.	**Re-evaluate molecular behavior after language-domain tuning.** If CC decreases, separate validity from constraint satisfaction among valid molecules before interpreting the loss.	Secs. 6.1–6.3; Fig. 4

Creativity in LLMs. Creativity research distinguishes convergent and divergent aspects of creative behavior (Guilford, 1967; Cropley, 2006; Runco & Jaeger, 2012). Recent LLM studies have evaluated creativity in open-ended language generation (Atmakuru et al., 2024; Chakrabarty et al., 2024; Lu et al., 2024), reasoning (Tian et al., 2024), and code (DeLorenzo et al., 2024; Lu et al., 2025). In molecular generation, exploration must be considered together with chemical validity and functional constraints. We therefore adopt the conceptual distinction between convergent and divergent creativity and instantiate it using standard molecular-generation measures. We further test whether preference tuning designed to improve creativity-related behavior in language produces corresponding gains in molecular generation.

3 Evaluation and Experimental Setup

Drawing on creativity theory (Guilford, 1967; Cropley, 2006; Runco & Jaeger, 2012), we evaluate molecular generation along two complementary dimensions (Figure 1). **Convergent creativity** (CC) captures validity and task success: generated molecules should be chemically valid and satisfy the requested constraints. **Divergent creativity** (DC) captures novelty, uniqueness, and structural diversity: generated molecules should retain exploration rather than collapse to a narrow set of structures.

3.1 Convergent Creativity

Let G be the total number of generations in a batch, V the number that RDKit (Landrum et al., 2013) parses as chemically valid molecules, and S the number that are valid and satisfy every task constraint. **Validity** is the fraction of all generations that are chemically valid, $\text{Validity} = V/G$, and **Success Rate** is the fraction of all generations that are valid and satisfy every task constraint, $\text{SR} = S/G$. Constraint satisfaction among valid molecules is $\text{SR} \mid \text{valid} = S/V$, which we later use to decompose CC in Section 6. Convergent creativity is the geometric mean of the two convergent metrics:

$$\text{CC} = \sqrt{\text{Validity} \times \text{SR}}. \tag{1}$$

A model therefore scores well only if it both produces valid molecules and satisfies the requested constraints.

3.2 Divergent Creativity

Novelty is the fraction of valid molecules absent from a fixed ZINC 250K reference corpus (Sterling & Irwin, 2015), $\text{Novelty} = N/V$; **Uniqueness** is the fraction of valid molecules that are not duplicates, $\text{Uniqueness} = U/V$; and **Diversity** is one minus the mean pairwise Tanimoto similarity over Morgan fingerprints (Rogers & Hahn, 2010) of the valid molecules. Because corpus-relative novelty does not capture reproduction of in-context examples, prompt copying is evaluated separately in Section 5.3. Divergent creativity is the geometric mean of the three:

$$\text{DC} = (\text{Novelty} \times \text{Uniqueness} \times \text{Diversity})^{1/3}. \tag{2}$$

We use geometric means so that a high value on one component cannot fully compensate for a low value on another; if any component is zero, the corresponding composite score is also zero.

3.3 Aggregation and Additional Summary Metrics

Each model–task result is averaged over five independent runs. Across 10 models and 15 evaluation tasks, this gives 150 model–task cells. The five LogP targets are five targets of one property rather than five distinct objectives: without family balancing they would receive one third of the weight across the 15 evaluation tasks. Cross-task summaries therefore average within the LogP family first and then across the 11 task families, giving 110 model–family observations, while analyses of task-level recurrence retain all 15 tasks and 150 cells. For secondary summaries we also report Overall Creativity, $\sqrt{\text{CC} \times \text{DC}}$, and the Fully Creative rate, the fraction E/G of all generations that are simultaneously successful, novel, and unique. Appendix B provides the complete metric definitions, implementation details, robustness analyses, and additional aggregation variants.

3.4 Models, Tasks, and Generation Settings

As shown in Figure 1 (Top), models generate SMILES from prompts specifying one or more molecular constraints; inference-time experiments additionally provide in-context molecular examples where stated.

Task description. We evaluate 15 evaluation tasks: 12 single-constraint tasks spanning physicochemical properties (QED, SA, and five LogP targets), ADMET properties (BBB and HIA), and biological activity (DRD2, JNK3, and GSK3β), together with three two-constraint tasks (QED+SA, BBB+QED, and HIA+QED). Task definitions and prompt templates are given in Appendix D.

Evaluation. To measure these properties, we use RDKit (Landrum et al., 2013) to calculate the LogP values of the generated molecules; TDC oracle functions (Huang et al., 2021) to obtain predictive scores for target activity on DRD2, JNK3, and GSK3β, as well as QED and SA; and models from ADMET-AI (Swanson et al., 2024) to obtain predictive scores for BBB and HIA. See Appendix C for details on the reliability of the predictors.

Key details. We conduct experiments across a diverse set of LLMs that vary in model size, series, and architecture, including LLaMA3-8B, LLaMA3-70B, LLaMA3.1-8B, LLaMA3.1-70B (Grattafiori et al., 2024), Qwen3-8B, Qwen3-32B (Yang et al., 2025), GPT-3.5, GPT-4o-mini (Hurst et al., 2024), GPT-5 (Singh et al., 2025), and DeepSeek-V3 (Liu et al., 2024). Provider and serving-route details are reported in Appendix P. Unless otherwise specified, the default sampling temperature is 1.0. Appendix G reports the task-balanced and category-balanced alternatives to the family-balanced weighting. Detailed results of all models across all evaluation tasks are provided in Appendix S.

4 Convergent and Divergent Creativity Are Negatively Associated

The negative association between convergent and divergent creativity appears in every evaluation task. Across all 15 tasks, models with higher CC also tend to have lower DC: every task-specific Pearson correlation is negative, with a median of $r = -0.577$ (Figure 2A). A pooled correlation could be produced by tasks differing in difficulty rather than by any relationship between the two dimensions, so we compute the correlation separately inside each task, holding the task fixed and comparing models within it. The 15 correlations are consistently negative (IQR $[-0.740, -0.327]$;

range $[-0.821, -0.074]$): the magnitude varies across tasks, but even the weakest association points the same way.

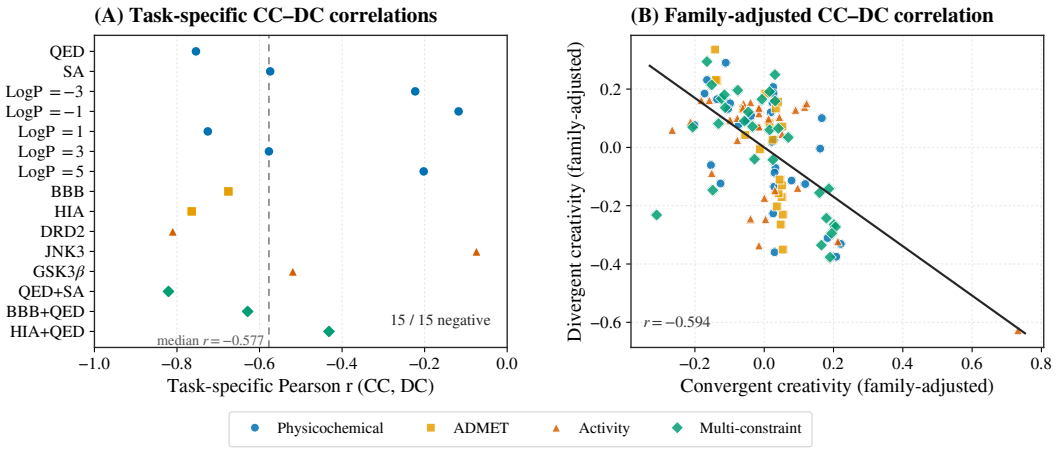


Figure 2: The CC–DC association recurs across evaluation tasks. **(A)** Task-specific Pearson correlations are negative in all 15 evaluation tasks; the dashed line marks the median ($r = -0.577$). **(B)** CC and DC remain negatively correlated after task-family adjustment ($r = -0.594$).

Task families, however, differ substantially in their typical CC and DC levels. To check whether these family-level differences explain the pattern, we remove each family's average CC and DC before comparing models. Because the five LogP targets are five targets of one property rather than five distinct objectives, we first combine them into a single family so that each of the 11 task families receives equal weight, giving 110 model–family observations. Over these observations the association is negative before any adjustment ($r = -0.556$) and remains so within every stratum: all four task categories (physicochemical -0.575, ADMET -0.695, activity -0.552, multi-constraint -0.565) and all four model families (LLaMA -0.582, GPT -0.487, Qwen -0.715, DeepSeek -0.833), with all 95% confidence intervals excluding zero (Appendix F). Subtracting each family's mean leaves the association negative as well ($r = -0.594$; Figure 2B).

The result also remains under fixed-effect models, which absorb these differences rather than averaging them away. Task-family fixed effects remove the average CC and DC level of each family, so the estimate rests on variation among models within a family; the CC–DC coefficient is $\beta = -0.847$ (family-block bootstrap 95% CI $[-1.268, -0.701]$). Adding model fixed effects further removes persistent model-level differences; the coefficient becomes smaller in magnitude but remains negative ($\beta = -0.455$; $[-0.661, -0.092]$). The sign is unchanged under model-block permutation, under leave-one-out refits (all 25 single-exclusion fits fall between -0.580 and -0.289), and under alternative task and category weighting (Appendices E and G).

The decomposition of DC shows what a lower value consists of. Repeating the task-family fixed-effect model with each divergent component as the outcome gives a negative slope for uniqueness ($\beta = -0.620$) and diversity ($\beta = -0.495$), while the novelty slope is indistinguishable from zero ($\beta = 0.0003$). Novelty carries little variation to explain: across the 150 model–task cells it spans only 0.965–1.000, with 134 of 150 at or above 0.99. In this benchmark, lower DC therefore primarily reflects more repeated and more structurally similar molecules, rather than lower novelty relative to the reference corpus.

Together, these analyses show that the negative relationship between convergent and divergent creativity is not limited to a particular task or task family, but recurs across the benchmark. Evaluating only constraint satisfaction or only exploration can therefore miss systematic changes in the other dimension.

Finding 1: Constraint satisfaction and molecular exploration should be evaluated together: across all 15 tasks, higher CC is associated with lower DC, and the relationship remains after accounting for task-family and model differences.

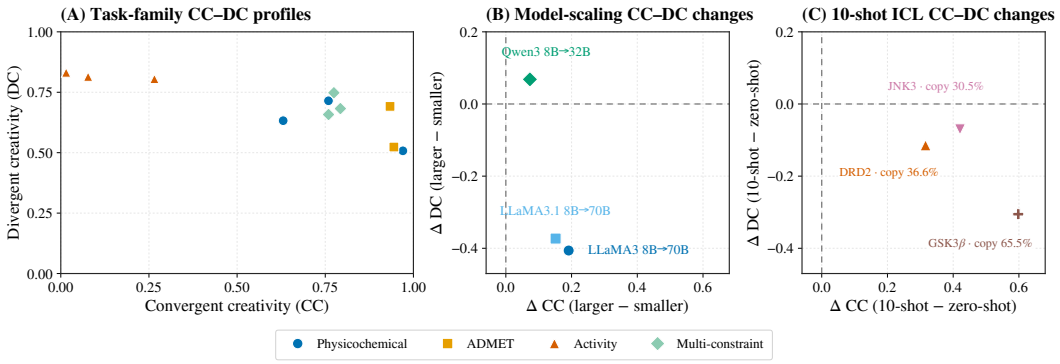


Figure 3: CC–DC profiles vary across task families, model families, and inference conditions. **(A)** Activity families combine low CC with high DC, while physicochemical and ADMET families reach higher CC. **(B)** Across size-varied checkpoints, LLaMA gains CC while losing DC, whereas Qwen3 improves both. **(C)** Ten-shot ICL increases activity-task CC while reducing DC; labels show exact-copy rates.

5 What Shapes Convergent and Divergent Creativity?

The negative CC–DC association of Section 4 holds across the benchmark, but tasks, models, and generation conditions still occupy very different regions of the CC–DC plane. We next examine three sources of this variation: task structure, model family across size-varied checkpoints, and inference-time conditioning (Figure 3).

5.1 Task-Side Factors

Task families occupy distinct regions of the CC–DC plane. Activity tasks occupy a low-CC/high-DC region, whereas physicochemical and ADMET tasks generally reach higher CC (Figure 3A). ADMET and activity runs differ in shape as well as position: ADMET runs concentrate in a narrow band at high CC, with almost all of their spread along the divergent axis, while activity runs sit near the left edge of the plane at near-zero CC with high DC (Appendix F). Multi-constraint families do not fall in the lowest-CC region.

In the tested compatible combinations (QED+SA, BBB+QED, and HIA+QED), adding a second constraint does not necessarily reduce success (Appendix I). A cross-property check on the two 8B LLaMA checkpoints is informative here: for LLaMA3-8B and LLaMA3.1-8B, prompting for SA alone reaches QED success of 0.522 and 0.552, above the 0.488 and 0.498 obtained when QED is requested directly, and prompting for BBB alone reaches 0.508 and 0.616. For these two checkpoints, SA- or BBB-conditioned generations already satisfy QED at substantial rates, indicating substantial overlap between molecules satisfying these properties. This helps explain why adding QED does not necessarily reduce success in the tested compatible combinations.

Task structure also shapes the form of control, not only its level. Across the five LogP targets the mode of the generated distribution shifts with the requested value (Spearman $\rho = 0.927$), but the distributions stay broadly dispersed around it, so numerical constraints give directional control rather than precise targeting (Appendix J). The activity profile is not specific to the oracle scorer: docking GPT-3.5 generations against 5-HT1B and ACM2 gives convergent creativity of 0.219 and 0.172 alongside uniqueness of 1.000 and 0.939, showing the same low-CC/high-uniqueness profile observed in the oracle-scored activity tasks (Appendix K).

5.2 Model-Side Factors

Across size-varied checkpoints, the LLaMA families gain CC while losing DC, whereas Qwen3 improves both. Figure 3B plots the change between the two evaluated sizes of each family; deltas are paired by task family. The larger LLaMA3 checkpoint moves right and down ($\Delta\text{CC} = +0.191$, 95% CI $[+0.118, +0.268]$; $\Delta\text{DC} = -0.406$, $[-0.492, -0.315]$; both $q = 0.004$), and the direction

holds in all 11 task families. LLaMA3.1 moves the same way ($\Delta\mathrm{CC} = +0.151$, $[+0.091, +0.215]$; $\Delta\mathrm{DC} = -0.373$, $[-0.457, -0.280]$; both $q = 0.004$) in 10 of 11.

Qwen3 8B→32B instead moves right and up, improving both dimensions in 8 of 11 families ($\Delta\mathrm{CC} = +0.073$, $[+0.023, +0.138]$, $q = 0.012$; $\Delta\mathrm{DC} = +0.068$, $[+0.028, +0.101]$, $q = 0.024$), and its overall creativity rises accordingly ($+0.096$, $[+0.039, +0.176]$, $q = 0.016$). The aggregate association of Section 4 is negative, but a checkpoint comparison can still raise both dimensions at once.

Two diagnostics clarify what accompanies the LLaMA CC gain. The Fully Creative rate, which counts molecules that are simultaneously successful, novel, and unique, falls from 0.556 to 0.205 between LLaMA3-8B and LLaMA3-70B and from 0.394 to 0.126 for LLaMA3.1, so the CC gain does not yield more molecules that satisfy all three conditions together. At the same time the larger checkpoints are more stable at the syntax level: the 8B checkpoints range from roughly 0.64 to 0.96 in validity across constraint settings, whereas the larger ones stay near-saturated (Appendices G and L).

5.3 Inference-Time Factors

Ten-shot ICL sharply increases activity-task success while reducing exploration. We compare GPT-3.5 zero-shot generation with ten-shot ICL using active examples on DRD2, JNK3, and GSK3β (Figure 3C). Success rate rises from 0.150 to 0.542 on DRD2, from 0.006 to 0.316 on JNK3, and from 0.032 to 0.698 on GSK3β, while DC falls on all three tasks (Appendix M). The DC decline primarily reflects lower uniqueness rather than lower novelty. Uniqueness falls from 0.957 to 0.688 on DRD2, from 0.951 to 0.763 on JNK3, and from 0.964 to 0.407 on GSK3β, while diversity moves materially only on GSK3β (0.875 to 0.643); novelty remains near ceiling throughout (≥ 0.993 in both conditions), consistent with the benchmark-wide pattern in Section 4. Validity also declines, most sharply on JNK3 (0.976 to 0.732), so the increase in activity success is accompanied by lower molecular validity.

Our standard novelty metric compares generated molecules against a fixed reference corpus, not against the ten in-context examples themselves. It therefore does not capture direct reuse of the demonstrations. We measure this separately using the exact-copy rate, defined as the fraction of valid generations whose canonical SMILES exactly matches one of the ten supplied examples. The rate is 0.000 under zero-shot prompting and rises to 0.366 on DRD2, 0.305 on JNK3, and 0.655 on GSK3β under ten-shot ICL. The increase in activity success is therefore accompanied by substantial direct reuse of the supplied examples. Allowing near-copies at Tanimoto ≥ 0.8 raises these rates only marginally, to 0.373, 0.310, and 0.701, so the resemblance to the demonstrations is mostly literal reproduction rather than a broad set of close analogues. Which examples are supplied also matters: replacing uniformly random examples with cluster medoids raises success rate on GSK3β ($+0.342$) and DRD2 ($+0.116$) while lowering DC (-0.219 and -0.078), while JNK3 moves in the opposite direction ($\Delta\mathrm{SR} = -0.124$, $\Delta\mathrm{DC} = +0.035$), with copying also highest on the two tasks where medoid selection increases success (Appendix M).

Other inference-time controls show task- and model-dependent effects. Decoding temperature moves the two dimensions continuously: sweeping GPT-3.5 on QED from 0.2 to 1.2 raises DC monotonically from 0.496 to 0.930, while CC peaks at 0.5 (0.841) and declines to 0.711. A similar pattern appears for LLaMA3-8B on SA, but the best overall setting sits elsewhere, peaking at temperature 1.25 (0.816). The best overall temperature differs between the two model–task pairs, suggesting that temperature should be tuned for the target model and task rather than transferred directly (Appendix N). Diversity-oriented prompting can also increase uniqueness without retraining. Conditioning GPT-3.5 on diverse anchors drawn from its own earlier outputs raises uniqueness from 0.892 to 0.978 on QED and from 0.763 to 0.961 on BBB, with overall creativity up $+0.022$ and $+0.035$ and CC essentially unchanged; on DRD2, where CC is already scarce, the same prompt cuts success rate from 0.150 to 0.074. The uniqueness gain replicates on a second checkpoint (0.819 to 0.917 on QED, 0.851 to 0.956 on BBB) while the net gain does not, since BBB success rate falls from 0.592 to 0.456 and overall creativity from 0.733 to 0.659 (Appendices O and P).

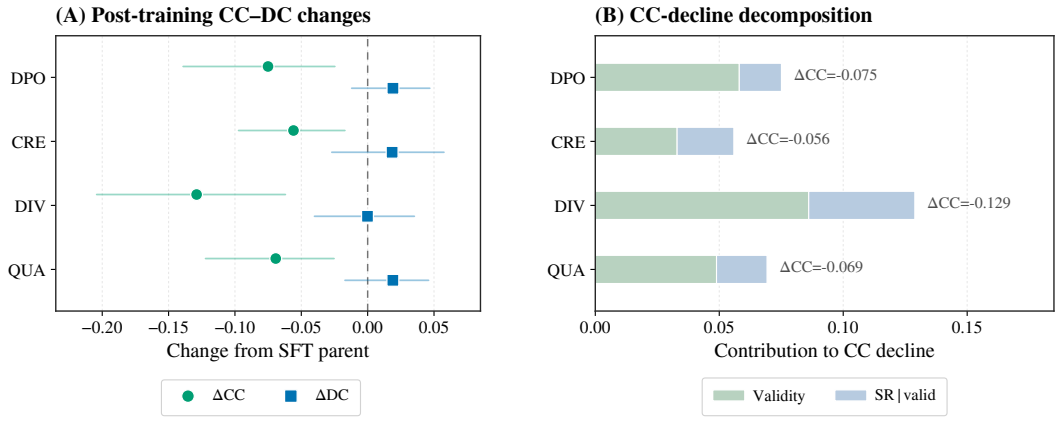


Figure 4: Preference tuning reduces molecular CC, while the creativity-oriented variants do not reproduce their targeted gains in molecular generation. **(A)** Mean changes in CC and DC relative to the shared SFT parent across 11 task families; error bars show 95% bootstrap CIs. **(B)** Each CC decline is decomposed into contributions from reduced validity and reduced constraint satisfaction among valid generations (SR | valid).

Finding 2: The CC–DC relationship is not fixed across tasks, models, or inference conditions. In particular, Qwen3 improves both dimensions across size-varied checkpoints, while ten-shot ICL increases activity success alongside reduced exploration and substantial reuse of demonstrations.

6 Does Creativity-Oriented Post-Training Transfer?

Creativity-oriented preference tuning is designed and evaluated in the language domain (Ismayilzada et al., 2025). We ask whether the same tuning also improves the corresponding aspects of molecular generation: if a model is tuned for novelty, diversity, or quality in language, does it show the same gain when generating molecules?

6.1 Controlled Same-Base Comparison

We test this question while holding the base model fixed. Four preference objectives are applied as LoRA adapters to the same CrPO-SFT Llama-3.1-8B parent and evaluated on all 15 evaluation tasks under the same generation and evaluation protocol. Plain DPO serves as the control, while three creativity-oriented CrPO variants target creativity/novelty/surprise (CRE), diversity (DIV), and quality (QUA). Thus, the comparison changes the preference objective while keeping the model family, scale, task set, decoding, and evaluator fixed.

Language-domain creativity tuning does not produce the targeted gains in molecular generation. All four preference-tuned variants have lower CC than the shared SFT parent, while their changes in DC are comparatively small (Figure 4A; exact values and significance in Table 14). None of the three creativity-oriented variants improves the molecular metric targeted by its language-domain objective. DIV, despite being tuned for diversity, shows essentially no DC improvement and the largest CC loss, while QUA has the largest DC gain among the three creativity-oriented variants but no improvement in CC. Plain DPO shows the same qualitative CC decline, so this pattern is not specific to the creativity-oriented objectives.

A second pattern is shared by all four adapters: molecular validity decreases after preference tuning.

6.2 Where Does the Performance Loss Come From?

Reduced validity accounts for most of the CC decline. Constraint success requires two things: the model must first generate a valid molecule, and that molecule must satisfy the requested property.

We therefore separate validity from constraint satisfaction among valid generations, measured as $\mathrm{SR} \mid \mathrm{valid}$.

Validity decreases for every preference-tuned variant relative to SFT: -0.089 for DPO, -0.048 for CRE, -0.117 for DIV, and -0.058 for QUA. The largest drop occurs for DIV. Constraint satisfaction among valid molecules also decreases overall, but the changes are less consistent across task families and do not survive multiplicity correction for any variant. In contrast, the validity decline remains significant for DPO ($q = 0.047$) and DIV ($q = 0.035$).

Because success requires validity, $\mathrm{SR} = \mathrm{validity} \times (\mathrm{SR} \mid \mathrm{valid})$. Together with the definition of CC, this gives $\mathrm{CC} = \mathrm{validity} \times \sqrt{\mathrm{SR} \mid \mathrm{valid}}$. We can therefore separate the observed CC decline into contributions from validity and from constraint satisfaction among valid molecules (Figure 4B; Appendix Q). Reduced validity accounts for 59–77% of the CC decline across the four variants, with reduced $\mathrm{SR} \mid \mathrm{valid}$ accounting for the remainder. When floor tasks are excluded, the validity contribution increases to 64–82%. The exact split also changes with task weighting (Appendix G), but validity remains the larger contributor.

6.3 Diagnostic Analysis of the Validity Drop

The validity decline is accompanied by longer and harder-to-synthesize molecular outputs. Median SMILES length increases from 36 characters under SFT to 44–54 across the four preference-tuned variants (all $q < 0.04$). Within every variant, longer generations are more likely to be invalid (point-biserial $r = -0.28$ to -0.40, $p < 10^{-27}$), and invalid generations are roughly three times longer than valid ones. The shift is also visible among valid molecules: synthetic accessibility increases for every variant (median ΔSA $+0.85$ to $+1.23$, all $q \leq 0.021$, across 13–14 of 15 tasks), accompanied by increases in lipophilicity and saturation.

The same ordering appears across the four preference-tuned variants: variants with longer outputs also have larger adapter updates, larger validity losses, and larger CC losses. With only four variants, however, the exact permutation test gives $p = 0.083$.

Format noncompliance accounts for only part of the decline. SFT follows the required JSON output format at a high rate, whereas every preference-tuned variant is less compliant. Restricting the analysis to format-compliant responses closes only 12–46% of the CC gap, and validity remains lower by 0.047–0.091 even among compliant responses. Additional adapter-geometry and likelihood analyses are reported in Appendix Q.

Finding 3: Language-domain creativity-oriented preference tuning does not reproduce its intended gains in molecular generation. In the controlled same-parent comparison, CC decreases across all four preference-tuned variants, with reduced validity accounting for most of the decline. The validity loss is accompanied by longer, more lipophilic, and harder-to-synthesize molecular outputs.

7 Conclusion

Molecular generation requires both constraint satisfaction and chemical-space exploration. We study these abilities as convergent and divergent creativity across 10 general-purpose LLMs and 15 molecular generation tasks. Higher convergent creativity is consistently associated with lower divergent creativity, although their balance varies across tasks, models, and inference-time conditions. Creativity-oriented preference tuning in language does not automatically transfer to molecular generation: in our controlled same-parent comparison, none of the three creativity-oriented variants reproduces its targeted molecular gain, while all four preference-tuned variants reduce convergent creativity, largely through lower validity. Based on these findings, we provide practical guidelines to evaluate both dimensions jointly, select models and inference strategies for the target task, monitor copying under in-context learning, and re-evaluate molecular behavior after cross-domain post-training. Overall, our study reveals recurring patterns in how constraint satisfaction and exploration behave together in LLM-based molecular generation and provides practical guidance for researchers working in this area.

AI USE STATEMENT

OpenAI Codex and Claude Code assisted with manuscript editing, consistency checking, code inspection, and figure-preparation workflows. AI-assisted outputs were checked against the manuscript sources, generated result files, executable build checks, and archived experiment records. The authors are responsible for the final text, claims, code, figures, and analyses.

ETHICS STATEMENT

This study uses public molecular data and computational predictors, with no new human or animal experiments. Predicted properties, activity scores, ADMET scores, and docking results do not establish clinical safety, efficacy, synthesizability in practice, or biological activity. Novelty is measured relative to a fixed reference corpus and should not be interpreted as chemical discovery or as absence from model pretraining data. Generated molecules may raise dual-use or safety concerns and require expert chemical and biological review before synthesis, testing, or deployment. Any released materials should retain the applicable data, model, and software license terms.

REPRODUCIBILITY STATEMENT

Sections 3 and 3.4 define the evaluation metrics, model set, task suite, generation protocol, and property predictors. Sections 4 to 6 report the main analyses, and Sections B to D, G, Q and S provide metric definitions, prompt templates, predictor details, robustness checks, detailed model–task results, and post-training diagnostics. The manuscript documents the metrics, prompts, model settings, robustness checks, and generated result summaries needed to audit the reported tables and figures. Release of code or generated result files is subject to the availability and license terms of the underlying models, APIs, datasets, and molecular-evaluation software.

REFERENCES

Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints. *arXiv preprint arXiv:2410.04197*, 2024.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–34, 2024.

Arthur Cropley. In praise of convergent thinking. *Creativity research journal*, 18(3):391–404, 2006.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

Matthew DeLorenzo, Vasudev Gohil, and Jeyavijayan Rajendran. Creativeval: Evaluating creativity of llm-based hardware code generation. In *2024 IEEE LLM Aided Design Workshop (LAD)*, pp. 1–5. IEEE, 2024.

Joseph A DiMasi, Lanna Feldman, Abraham Seckler, and Andrew Wilson. Trends in risks associated with new drug development: success rates for investigational drugs. *Clinical Pharmacology & Therapeutics*, 87(3):272–277, 2010.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in neural information processing systems*, 35: 21342–21357, 2022.

Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Joy P Guilford. Creativity: Yesterday, today and tomorrow. *The Journal of Creative Behavior*, 1(1): 3–14, 1967.

Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Xiuyuan Hu, Guoqing Liu, Quanming Yao, Yang Zhao, and Hao Zhang. Hamiltonian diversity: effectively measuring molecular diversity by shortest hamiltonian circuits. *Journal of Cheminformatics*, 16(1):94, 2024.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Mete Ismayilzada, Antonio Laverghetta, Simone A Luchini, RN Patel, Antoine Bosselut, Lonneke van der Plas, and Roger E Beaty. Creative preference optimization. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9580–9609, 2025.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International conference on machine learning*, pp. 4839–4848. PMLR, 2020a.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pp. 4849–4859. PMLR, 2020b.

Jaehyeong Jo, Dongki Kim, and Sung Ju Hwang. Graph generation with diffusion mixture. *arXiv preprint arXiv:2302.03596*, 2023.

Greg Landrum et al. Rdkit documentation. *Release*, 1(1-79):4, 2013.

Seul Lee, Karsten Kreis, Srimukh P Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Arash Vahdat, and Weili Nie. Molecule generation with fragment retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:132463–132490, 2024.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Paliwal, Weili Nie, and Arash Vahdat. Genmol: A drug discovery generalist with discrete diffusion. *arXiv preprint arXiv:2501.06158*, 2025.

Bingxuan Li, Yiwei Wang, Tao Meng, Kai-Wei Chang, and Nanyun Peng. Control large language models via divide and conquer. *arXiv preprint arXiv:2410.04628*, 2024a.

Jiatong Li, Junxian Li, Yunqing Liu, Dongzhan Zhou, and Qing Li. Tomg-bench: Evaluating llms on text-based open molecule generation. *arXiv preprint arXiv:2412.14642*, 2024b.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE transactions on knowledge and data engineering*, 36(11):6071–6083, 2024c.

Yibo Li, Liangren Zhang, and Zhenming Liu. Multi-objective de novo drug design with conditional graph generative model. *Journal of cheminformatics*, 10(1):33, 2018.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Yiping Liu, Jiahao Yang, Xuanbai Ren, Zhang Xinyi, Yuansheng Liu, Bosheng Song, Xiangxiang Zeng, and Hisao Ishibuchi. Multi-objective molecular design through learning latent pareto set. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19006–19014, 2025.

Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. Reinvent 4: Modern ai-driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, et al. Ai as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text. *arXiv preprint arXiv:2410.04265*, 2024.

Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2776–2794, 2025.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):48, 2017.

Tiago Pereira, Maryam Abbasi, Jose Luis Oliveira, Bernardete Ribeiro, and Joel Arrais. Optimizing blood–brain barrier permeation through deep reinforcement learning for de novo drug design. *Bioinformatics*, 37(Supplement_1):i84–i92, 2021.

Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:565644, 2020.

Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev. Molecularrnn: Generating realistic molecular graphs with optimized properties. *arXiv preprint arXiv:1905.13372*, 2019.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382*, 2020.

Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International conference on artificial neural networks*, pp. 412–422. Springer, 2018.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Yuxuan Song, Jingjing Gong, Minkai Xu, Ziyao Cao, Yanyan Lan, Stefano Ermon, Hao Zhou, and Wei-Ying Ma. Equivariant flow matching with hybrid probability transport for 3d molecule generation. *Advances in Neural Information Processing Systems*, 36:549–568, 2023.

Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of chemical information and modeling*, 55(11):2324–2337, 2015.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. Admet-ai: a machine learning admet platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024.

Wen Tao, Jing Tang, Alvin Chan, Bryan Hooi, Baolong Bi, Nanyun Peng, Yuansheng Liu, and Yiwei Wang. How to make large language models generate 100% valid molecules? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26576–26591, 2025.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5303–5324, 2024.

Bailin Wang, Zi Wang, Xuezhi Wang, Yuan Cao, Rif A Saurous, and Yoon Kim. Grammar prompting for domain-specific language generation with large language models. *Advances in Neural Information Processing Systems*, 36:65030–65055, 2023.

Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules. *arXiv preprint arXiv:2112.12542*, 2021.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

Peng Zhou, Jianmin Wang, Chunyan Li, Zixu Wang, Yiping Liu, Siqi Sun, Jianxin Lin, Leyi Wei, Xibao Cai, Houtim Lai, et al. Instruction multi-constraint molecular generation using a teacher-student large language model. *BMC biology*, 23(1):105, 2025.

APPENDIX CONTENTS

A Limitations **15**

B Evaluation Metrics **16**

B.1 Convergent Creativity Metrics 16

B.2 Divergent Creativity Metrics 16

B.3 Composite Metrics 17

B.4 Elite Subset Metric 17

B.5 Justification for Geometric Mean Aggregation 17

B.6 Implementation Details 18

C Reliability of Predictors **18**

D Prompt Templates **18**

E Task-Controlled Trade-off Diagnostics **19**

F Aggregate Task Profiles and Stratified Association **19**

G Methodological Robustness Checks **20**

H Alternative Diversity Definitions **22**

I Compatible Constraints **23**

J Numerical Constraints **23**

J.1 Additional Visualizations 24

K Docking-Based Objectives **25**

L Additional Model-Side Analyses **25**

M In-Context Learning Results **26**

N Temperature Sensitivity **27**

O Anchor-Conditioned Prompting **28**

P Cross-Model Validation of the Prompt Intervention **29**

Q Post-training Transfer Diagnostics **30**

R Specialist Molecular Generators: External Context **31**

R.1 Published Numbers for a Wider Set of Specialist Methods 31

S Detailed Results **33**

A LIMITATIONS

Despite the systematic evaluation presented in this work, several limitations remain. Being explicit about these issues is important, both for correctly interpreting the results and for guiding future research on LLM-based molecular generation.

Ambiguity of creativity metrics. A first limitation lies in how creativity is operationalized. Although we decompose creativity into convergent and divergent components using measurable indicators, these metrics are still proxies rather than direct measurements of chemical creativity. Different choices of metrics can lead to different quantitative conclusions. As a result, current creativity metrics capture only partial aspects of what chemists would intuitively regard as creative molecular design.

Dependence on the training corpus. LLM behavior is strongly shaped by the chemical patterns present in their training data. Many global properties, such as drug-likeness or simple physicochemical trends, can be inferred from frequent motifs and correlations in large text or SMILES corpora. In contrast, protein-specific activity often depends on subtle structure–activity relationships that are sparsely represented or inconsistently annotated in public data.

Novelty is measured against a reference corpus, not a training set. Novelty is defined relative to a reference corpus, and no finite corpus represents the full space of known chemistry. Ours is ZINC 250K, fixed across every experiment. We do not have access to any evaluated model's actual pretraining data, so a molecule counted as novel may still have been seen during training, and novelty here should be read as "absent from a standard corpus" rather than as discovery. Recomputing against ChEMBL, roughly six times larger and pharmacologically enriched, lowers novelty in all 150 model–task cells, by more than 0.3 in 31 of them, and drops 30 of the 134 cells scoring at least 0.99 against ZINC below 0.70 (Appendix G). High apparent novelty against a small corpus can therefore mask substantial overlap with known bioactive chemistry; GPT-5 on DRD2 is the sharpest case, scoring 1.000 against ZINC and 0.287 against ChEMBL while emitting four known actives repeatedly (Table 27). Because the two corpora order cells almost independently ($\rho = -0.02$), neither absolute novelty values nor novelty-based model comparisons should be read as evidence of chemical discovery. Our conclusions rest on neither: divergent creativity is reproduced almost exactly without the novelty component at all ($\rho = 0.9999$), so it is carried by uniqueness and diversity, which are computed entirely within a generated batch and are immune to the concern. This is why we report them alongside novelty rather than folding the dimension into a single number, and why prompt-level reproduction is measured separately as a copy rate (Appendix M) rather than by widening the corpus.

Success rates depend on threshold conventions. Constraint satisfaction requires committing to a numerical cutoff. As documented in Appendix G, model rankings are sensitive to this choice: sweeping the QED threshold shifts individual models by up to nine rank positions. The direction of the convergent–divergent trade-off and the absolute magnitude of success rates are stable, but close rankings between models with similar success rates should not be treated as threshold-independent.

Oracle-based rather than experimental evaluation. Activity and ADMET properties are scored with computational predictors rather than measured experimentally (Appendix C). These predictors are widely used and well validated, but they are models themselves, and a molecule that satisfies a predictor has not thereby been shown to satisfy the underlying biology.

The trade-off is a property of the generators we evaluate. Our central finding characterizes how a particular class of systems—general-purpose LLMs sampling under prompt conditioning—distributes its behavior across the two creativity dimensions. It is not a statement that constrained molecular generation must involve this trade-off. Generators built with explicit diversity-maintaining components occupy the plane differently, and the trade-off should be read as describing default generative behavior rather than an inherent limit.

Scope of the transfer diagnostics. The controlled comparison in Section 6 holds the base model fixed and covers the full task suite, but it rests on four preference objectives, one training run each, from a single model family, with training seed not held constant across variants (the adapter analysis finds no seed artifact in the weights, which is reassuring, but a single run per objective still carries that run's training noise). Every cross-variant claim therefore has $n = 4$: the rank concordances between update norm, output length, validity and ΔCC are perfect but have an exact permutation p of 0.083, which is consistent evidence rather than a quantitative law. Most importantly, the link from output elaboration to validity loss is *correlational*: length predicts invalidity within every variant and

Δlength rank-orders with the damage, but we did not run a length-matched or decode-capped test, so the direction of causation is not established. We describe elaboration as the diagnostic pattern most consistent with the evidence and do not claim it is demonstrated.

Limited replication of the inference-time factors. The temperature and prompt-intervention analyses each cover two settings rather than one, which is enough to separate what replicates from what does not but not enough to establish either as general. For temperature, the direction reproduces across both model–task pairs while the location of the optimum does not (Appendix N). For the prompt intervention, the uniqueness gain reproduces on a second model while the net creativity gain does not, and the second model is an SFT derivative evaluated at 25 samples per run with sequentially dependent C2 samples, so it constrains the claim rather than settling it (Appendix P). Neither factor should be generalized across model families without further replication.

B Evaluation Metrics

This section gives the formal definition of every quantity used in the paper, the aggregation rules that combine them, and the implementation details needed to reproduce them.

Table 2: Evaluation framework for LLM creativity in molecule generation. Convergent metrics assess constraint satisfaction, divergent metrics assess exploration, and composite metrics balance both dimensions. $\mathrm{GM}(\cdot)$ denotes the geometric mean.

Metric Category	Metric Name	Definition / Purpose	Role in Framework
Convergent	Validity	Are the molecules chemically real?	Foundational Check
	Success Rate	Do they meet the prompt's constraints?	Constraint Check
Divergent	Novelty	Are they absent from the reference corpus?	Reference-corpus novelty check
	Uniqueness	Are they different from each other?	Variety Check
	Diversity	How structurally different is the batch?	Exploration Check
Composite	Convergent Creativity	$\mathrm{GM}(\text{Validity}, \text{Success Rate})$	Convergent Creativity Score
	Divergent Creativity	$\mathrm{GM}(\text{Novelty}, \text{Uniqueness}, \text{Diversity})$	Divergent Creativity Score
	Overall Creativity	$\mathrm{GM}(\text{Convergent Creativity}, \text{Divergent Creativity})$	Overall Creativity Score
Elite Subset	Fully Creative	Proportion simultaneously Novel, Unique, & Successful.	Joint success / novelty / uniqueness diagnostic

We provide formal definitions for all metrics used to evaluate LLM creativity in molecular generation. Let G denote the total number of generated molecules and R denote the reference corpus.

B.1 Convergent Creativity Metrics

Validity. Measures whether generated SMILES strings correspond to chemically valid molecular structures. We use RDKit (Landrum et al., 2013) to verify the validity. Let G denote the total number of generated molecules and V denote the number of molecules successfully parsed by RDKit:

$$\text{Validity} = \frac{V}{G}. \tag{3}$$

Success Rate (SR). Measures the proportion of generated molecules that satisfy all task-specific constraints. Let S denote the number of valid molecules that meet all specified constraints. For multi-constraint tasks, a molecule is considered successful only if it satisfies *all* constraints simultaneously:

$$\text{Success Rate} = \frac{S}{G}. \tag{4}$$

B.2 Divergent Creativity Metrics

Novelty. Measures the fraction of generated molecules that do not appear in the reference corpus R. Following prior work (Gao et al., 2022; Li et al., 2024b), we set R to ZINC 250K (Sterling & Irwin, 2015), canonicalized, giving $249{,}455$ molecules. R is fixed across every model, task and experimental condition in this paper; nothing used to build a condition—in particular the activity datasets supplying in-context examples—enters it. Let N denote the number of valid molecules not present in R:

$$\text{Novelty} = \frac{N}{V}. \tag{5}$$

Uniqueness. Measures the proportion of non-duplicate molecules among valid generated molecules. This metric ensures the model does not collapse to producing only a few repetitive structures. Let U denote the number of unique molecules (after removing duplicates):

$$\text{Uniqueness} = \frac{U}{V}. \tag{6}$$

Diversity. Quantifies the average structural dissimilarity among generated valid molecules. We compute pairwise Tanimoto similarity using Morgan fingerprints (Rogers & Hahn, 2010) (radius 3, 2048 bits), then aggregate across all pairs. Let $V = \{m_1, \ldots, m_n\}$ denote the set of valid molecules, and $\text{sim}(m_i, m_j)$ denote the Tanimoto similarity between molecules m_i and m_j:

$$\text{Diversity} = 1 - \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \text{sim}(m_i, m_j). \tag{7}$$

Higher diversity indicates broader exploration of chemical space.

B.3 Composite Metrics

Convergent Creativity. Aggregates validity and success rate using their geometric mean:

$$\text{Convergent Creativity} = \sqrt{\text{Validity} \times \text{Success Rate}}. \tag{8}$$

Divergent Creativity. Aggregates novelty, uniqueness, and diversity using their geometric mean:

$$\text{Divergent Creativity} = \left(\text{Novelty} \times \text{Uniqueness} \times \text{Diversity}\right)^{1/3}. \tag{9}$$

Overall Creativity. Combines convergent and divergent creativity:

$$\text{Overall Creativity} = \sqrt{\text{Convergent Creativity} \times \text{Divergent Creativity}}. \tag{10}$$

The geometric mean ensures that both dimensions must be reasonably high to achieve a strong overall score, preventing models from excelling on only one dimension.

B.4 Elite Subset Metric

Fully Creative. Measures the proportion of molecules that are simultaneously successful (satisfy all task constraints), novel (not in the reference corpus), and unique (non-duplicate within the batch). Let E denote the number of molecules satisfying all three conditions:

$$\text{Fully Creative} = \frac{E}{G}. \tag{11}$$

This metric identifies generated molecules that simultaneously satisfy the task, are unique within the batch, and are absent from the chosen reference corpus.

B.5 Justification for Geometric Mean Aggregation

We aggregate component metrics with the geometric mean rather than the arithmetic mean because two of its properties match what a creativity score must express. For metrics $x_1, \ldots, x_n \in [0, 1]$,

$$\text{GM}(x_1, \ldots, x_n) = \left(\prod_{i=1}^{n} x_i\right)^{1/n}. \tag{12}$$

Property 1 (penalizes imbalance). $\text{GM}(x_1, \ldots, x_n) \le \text{AM}(x_1, \ldots, x_n)$, with equality if and only if $x_1 = \cdots = x_n$. A model therefore cannot compensate for a weak dimension by excelling on another, which is the behavior we want when convergent and divergent creativity are in tension.

Property 2 (zero-failure). If any component is zero the aggregate is zero. This matters concretely: a model emitting only invalid molecules (Validity $= 0$) or only duplicates (Uniqueness $= 0$) should receive no credit whatever its remaining scores.

The two properties together rule out the degenerate cases that a single averaged score would otherwise hide. A model with Validity $= 1.0$ and SR $= 0.0$ produces chemically well-formed molecules that satisfy no constraint; the geometric mean assigns convergent creativity 0.0, whereas the arithmetic mean would report 0.5 and place it mid-field. The same argument applies to the divergent components, where Novelty $= 0$ (complete overlap between valid outputs and the reference corpus) is correctly scored as no corpus-relative novelty. Table 27 reports a real instance of this failure mode rather than a hypothetical one.

We stress that the composite scores are a convenience for ranking within a fixed dimension, not a substitute for reporting the two dimensions separately. The central argument of this paper is that convergent and divergent creativity must be read jointly (Figure 5); accordingly we do not rank models by Overall Creativity anywhere in the main text.

Empirical non-redundancy. The recomputed correlations of Table 5 support aggregating within each dimension but not across them. Convergent and divergent creativity are negatively associated in every task category and model family ($r = -0.37$ to -0.80), providing the descriptive basis for the task-controlled analysis in Section 4. Within the divergent dimension the components are *not* interchangeable: uniqueness and diversity move together closely ($r = 0.802$) while novelty is only weakly related to diversity ($r = 0.161$, $p = 0.049$). This is an argument for keeping all three components rather than collapsing them, since novelty carries information the other two do not.

B.6 Implementation Details

For Tanimoto similarity calculations, we use RDKit's Morgan fingerprint implementation with default parameters (radius=3, nBits=2048). Reference dataset ZINC 250K (Sterling & Irwin, 2015) is preprocessed to canonical SMILES representation to ensure consistent comparison.

C Reliability of Predictors

Following existing work (Jin et al., 2020b;a; Lee et al., 2025), we use established property predictors to evaluate generated molecules, as real experimental measurements of activity and ADMET properties are prohibitively expensive for large-scale evaluation conducted in our work. Using validated predictors enables consistent and scalable comparison across methods.

The predictors used in our work are widely adopted in the community (Huang et al., 2021; Gao et al., 2022; Lee et al., 2024) and have demonstrated strong predictive performance. For example, prior work reports AUC values of 0.983 for JNK3 and 0.984 for GSK3 (Li et al., 2018), and above 0.85 for HIA and BBB predictors (Swanson et al., 2024). These results support the reliability of the models used in our evaluation setting.

D Prompt Templates

Definition and prompt template for each single-constraint task family. `[VALUE]` denotes one of the five LogP target values supplied per task instance.

Table 3: Task definitions and prompts.

Task	Property type	Constraint	Prompt
QED	Physicochemical	QED ≥ 0.6	The molecule has a high QED score.
SA	Physicochemical	SA ≤ 4	The molecule has good synthetic accessibility.
LogP	Physicochemical	LogP $\in \{-3, -1, 1, 3, 5\}$	The molecule has a LogP value of [VALUE].
BBB	ADMET	BBB ≥ 0.5	The molecule can pass through the blood-brain barrier.
HIA	ADMET	HIA ≥ 0.5	The molecule can be absorbed by the human intestine.
DRD2	Activity	DRD2 ≥ 0.5	The molecule can bind to DRD2.
JNK3	Activity	JNK3 ≥ 0.5	The molecule can bind to JNK3.
GSK3β	Activity	GSK3$\beta \geq 0.5$	The molecule can bind to GSK3β.

E TASK-CONTROLLED TRADE-OFF DIAGNOSTICS

All analyses in this section are computed at the evaluation-task level: one model–task cell after averaging five sampling runs. They support the within-task recurrence claim, so the task rather than the task family is the appropriate unit; the family-balanced counterparts of the fixed-effect estimates reported in Section 4 are given in Table 6. The verified table contains 10 models, 15 evaluation tasks, and 150 cells. No run-level observations are treated as independent, and no molecule generation, API call, or oracle query is performed.

Table 4: Task-level diagnostics for the CC–DC association, retaining all 15 evaluation tasks. Bootstrap intervals use 10,000 resamples. The task-FE permutation uses one model-label permutation shared across all tasks, preserving each model's cross-task structure while breaking the CC–DC pairing.

Analysis	Estimate	95% interval	Check	Interpretation
Pooled Pearson r	-0.496	$[-0.608, -0.364]$	$p = 1.12\times10^{-10}$	Descriptive baseline
Between-task Pearson r	-0.750	$[-0.912, -0.386]$	$p = 0.0013$	Task means differ systematically
Task-demeaned slope	-0.520	$[-0.686, -0.353]$	$r = -0.452$	Negative within-task association
Task FE slope	-0.520	$[-0.878, -0.307]$	HC3 $p = 1.27\times10^{-4}$; perm. $p = 0.0014$	Task-level controlled estimate
Within-task standardized slope	-0.527	$[-0.690, -0.363]$	HC3 $p = 2.14\times10^{-9}$	Same pattern on standardized scale
Task+model FE slope	-0.259	$[-0.517, -0.093]$	HC3 $p = 0.0166$; clustered $p = 0.028$	Negative after model baselines

Within-task recurrence. All 15 evaluation tasks are mathematically estimable, and all 15 have negative within-task slopes. The median within-task Pearson correlation is -0.577 (IQR $[-0.740, -0.327]$; range $[-0.821, -0.074]$), and the median slope is -0.775 (IQR $[-1.471, -0.567]$; range $[-2.119, -0.086]$). Bootstrapping over tasks gives a 95% interval of $[-0.725, -0.223]$ for the median Pearson correlation and $[-1.592, -0.547]$ for the median slope. The sign pattern is therefore recurring, while the magnitude varies substantially across tasks and task categories.

Leave-one-out stability. The two-way fixed-effect coefficient remains negative after leaving out any single evaluation task, model, or model family. Across leave-one-task fits the coefficient ranges from -0.374 to -0.205; across leave-one-model fits it ranges from -0.306 to -0.190. The largest influence is the LogP$= -3$ task, whose removal makes the coefficient more negative (-0.374), so no single task or model is needed to produce the negative controlled association.

Component-level behavior. Repeating the task-FE model with each divergent component as the outcome shows that the controlled association is carried by uniqueness and diversity, not by novelty. The slope is negative for uniqueness ($\beta = -0.620$, HC3 95% CI $[-0.924, -0.316]$, $p = 9.17\times10^{-5}$) and diversity ($\beta = -0.495$, $[-0.776, -0.214]$, $p = 6.70\times10^{-4}$), while novelty is near zero ($\beta = 0.0003$, $[-0.005, 0.006]$, $p = 0.902$). This diagnostic does not redefine divergent creativity; it indicates which component carries the task-controlled CC–DC relationship.

F AGGREGATE TASK PROFILES AND STRATIFIED ASSOCIATION

This section provides additional evidence for Section 4: the convergent–divergent association within each task category and model family, and the task profiles that motivate reading the two dimensions jointly.

Task categories occupy different parts of the CC–DC plane. On physicochemical and ADMET constraints (e.g., QED, SA, BBB, HIA), general-purpose LLMs reach high zero-shot success rates: ADMET runs concentrate in a narrow band at high convergent creativity, and the spread is almost entirely along the divergent axis. For target-specific activity tasks (DRD2, JNK3, GSK3β), the mass of runs collapses to the left edge—near-zero convergent creativity combined with high divergent creativity.

The same aggregate pattern holds for the individual convergent metrics: validity and success rate are each negatively correlated with diversity ($r = -0.466$ and $r = -0.400$). The raw property score is *not* uniformly opposed to diversity: pooled across tasks the correlation is indistinguishable from zero ($r = 0.03$, $p = 0.75$), and it is strongly negative only for QED ($r = -0.77$), QED+SA ($r = -0.93$), and DRD2 ($r = -0.86$), while BBB is positive ($r = +0.31$) and HIA, JNK3, GSK3β, and most LogP targets are near zero. The trade-off is therefore a property of the aggregate creativity dimensions, not of every constraint taken individually.

Table 5: Descriptive aggregate correlation between convergent and divergent creativity, overall and within strata. The unit of observation is the model–family observation after five-run model–task means are averaged within task families.

Stratum	**Group**	r **(CC, DC)**	**95% CI**
Overall	All (n=110)	-0.556	$[-0.67, -0.41]$
Task category	Physicochemical	-0.575	$[-0.78, -0.27]$
	ADMET	-0.695	$[-0.87, -0.36]$
	Activity	-0.552	$[-0.76, -0.24]$
	Multi-constraint	-0.565	$[-0.77, -0.26]$
Model family	LLaMA	-0.582	$[-0.75, -0.34]$
	GPT	-0.487	$[-0.71, -0.17]$
	Qwen	-0.715	$[-0.87, -0.42]$
	DeepSeek	-0.833	$[-0.96, -0.47]$

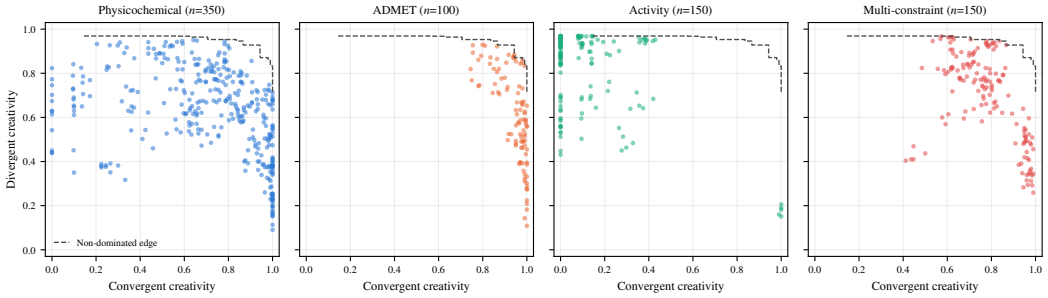


Figure 5: Every evaluated run placed in the convergent–divergent creativity plane, grouped by task category. The dashed staircase summarizes the non-dominated edge of the pooled run-level cloud and is descriptive rather than a task-controlled estimate. Task categories occupy visibly different parts of the plane.

G Methodological Robustness Checks

Every check in this section reuses the existing generations: no molecule was generated, and no LLM, API, oracle, or ADMET predictor was re-queried. The checks cover 750 runs, 150 model–task cells, 75,000 generations, and 70,183 valid instances. All 1,679 input files were SHA256-hashed before and after execution and verified unchanged, and the pipeline's 85 automated acceptance checks all pass.

Continuous versus binary novelty. Our novelty metric asks whether a generated molecule appears in the reference corpus, which is a hard threshold on a continuous quantity. We recomputed novelty as $1 - \max$ Tanimoto similarity against the same fixed corpus and rebuilt divergent creativity from it. The two definitions rank model–task cells almost identically (Spearman $\rho = 0.986$, $n = 150$; per-task ρ between 0.85 and 1.00; model-average ranking $\rho = 0.976$), so the conclusions in the main text are unaffected by the binary definition. We retain the binary form for comparability with prior work and report the continuous form here. One caveat applies to the continuous form: a distance of zero indicates identical Morgan fingerprints, which fingerprint collisions can produce without exact structural identity, so it is not by itself proof of membership in the reference corpus.

Success-rate threshold sensitivity. Success rate depends on a threshold that is conventional rather than derived. Sweeping the QED threshold from 0.40 to 0.80 while holding all other constraints fixed, model rankings are *not* stable: against the 0.60 threshold used throughout the paper, the rank correlation falls as low as $\rho = 0.067$, top-3 overlap as low as 0.33, and individual models shift by up to 9 rank positions. Absolute success rates and the direction of the trade-off are unaffected, but *fine-grained rankings between models with similar success rates should not be read as threshold-*

independent. The model ordering is unstable across the whole 0.40–0.80 scan rather than at isolated points, so this is a genuine limitation of any threshold-based success criterion, ours included.

Diversity and sample size. Because diversity is computed over the valid molecules in a batch, a model with low validity contributes fewer molecules and could in principle obtain an inflated diversity score. Recomputing diversity on a fixed sample of 50 valid molecules, averaged over 100 resamples, reproduces the original values almost exactly (Spearman $\rho = 0.9999$, maximum absolute difference 0.0053). Low validity therefore affects the uncertainty of the diversity estimate, not its expected value. One of the 750 runs (Qwen3-8B on DRD2, run 4) yielded only 43 valid molecules and could not be resampled at $n = 50$; it is recorded as an explicit insufficient-sample case rather than silently padded or dropped.

Fingerprint choice. Recomputing diversity with MACCS keys instead of Morgan fingerprints preserves the overall picture (Pearson $r = 0.930$, Spearman $\rho = 0.825$; model-average ranking $\rho = 0.939$), but the agreement is weaker than for the other checks. Overall and model-average trends are robust; absolute diversity values and some within-model cross-task patterns should be treated as fingerprint-dependent.

Paired statistics. All deltas reported in Sections 5.2 and 6 use the task *family* as the unit of observation: the five runs are averaged within an evaluation task, the tasks are averaged within their family, and the 11 families are then paired. Confidence intervals come from 10,000 bootstrap resamples, p-values from exact paired permutation tests, and multiplicity is controlled with Benjamini–Hochberg across all confirmatory comparisons. Of the 45 confirmatory tests, 25 are significant at $q < 0.05$; the remainder we describe as trends rather than effects. Comparisons with only six tasks are labelled exploratory and are not FDR-corrected, since the correction is uninformative at that scale—this is why no CRE – SFT comparison in Section 6 is reported as significant.

Weighting of evaluation tasks. The benchmark runs 15 evaluation tasks spanning 11 task families; the five LogP targets are five targets of one property rather than five distinct objectives. Weighting tasks equally would therefore give LogP roughly a third of the cross-task weight, so every aggregate in this paper averages within a family first and then across the 11 families. Table 6 reports the convergent–divergent association under that choice and under the two alternatives.

Table 6: Convergent–divergent association under three weightings of the evaluation tasks. The 11-family row is the specification used throughout the paper. Brackets give ordinary model-based 95% confidence intervals for the weighting comparison; Section 4 reports family-block bootstrap CIs for the primary 11-family fixed-effect estimates ($[-1.268, -0.701]$ and $[-0.661, -0.092]$). The category-balanced row admits no model fixed effect at $n = 40$.

Weighting	Pooled r	Task FE β	Task + model FE β
15 evaluation tasks ($n = 150$)	-0.496	-0.520 $[-0.695, -0.345]$	-0.259 $[-0.398, -0.121]$
11 task families ($n = 110$)	$\mathbf{-0.556}$	$\mathbf{-0.847}$ $[-1.077, -0.617]$	$\mathbf{-0.455}$ $[-0.613, -0.298]$
4 task categories ($n = 40$)	-0.572	-1.140 $[-1.586, -0.694]$	—

The choice is not neutral, and its effects run in both directions. Family weighting lowers the physicochemical share of the total weight from 46.7% to 27.3% and raises activity and multi-constraint to 27.3% each, so the harder activity families count for more. The convergent–divergent association strengthens accordingly (pooled r from -0.496 to -0.556; both fixed-effects slopes steepen), and of the 45 confirmatory paired tests 25 rather than 20 survive Benjamini–Hochberg correction, because averaging five correlated LogP tasks into one observation removes noise rather than adding it.

The transfer decomposition moves the other way. The validity component accounts for 59–77% of the convergent loss under the primary family weighting, compared with 72–94% under task weighting, because $\Delta(\mathrm{SR} \mid \mathrm{valid})$ grows for every variant (-0.042 to -0.124, against -0.015 to -0.082). Validity remains the larger of the two decomposition terms for all four variants, so the qualitative attribution in Section 6 is unchanged, but it is a majority rather than a near-totality, and the split depends on this weighting choice. Scaling is insensitive by comparison: the LLaMA families lose more divergent creativity, not less ($-0.326 \to -0.406$ and $-0.310 \to -0.373$), and Qwen3 remains positive on both dimensions.

One result in Section 4 is deliberately left at the task level: that all 15 within-task slopes are negative. Its claim is that the association reappears *inside* each individual evaluation task, so the task, not the family, is the unit that claim is about. The fixed-effects models in the same section are family-balanced like every other aggregate in the paper.

Inference and stability for the family-balanced fixed-effect estimates. The task-family fixed-effect slope reported in Section 4 ($\beta = -0.847$) has HC3 $p = 5.8\times10^{-9}$ and a model-block permutation $p = 0.0010$, using the same model-label permutation across families, which preserves each model's cross-family structure while breaking the CC–DC pairing. Leave-one-out refits of the two-way model do not reverse the sign: across all 25 single-exclusion refits the coefficient stays between -0.580 and -0.289, with the largest change after omitting the LLaMA family.

Does divergent creativity depend on the novelty component at all? Novelty sits near ceiling throughout: across the 150 model–task cells it spans only 0.965–1.000, with 134 of 150 at or above 0.99. Recomputing divergent creativity without it, as $\mathrm{GM}(\text{uniqueness}, \text{diversity})$, reproduces the reported ordering almost exactly (Spearman $\rho = 0.9999$ across cells, 0.988 across model averages; mean absolute difference 0.098, entirely a level shift). Every conclusion in this paper that rests on divergent creativity is therefore carried by uniqueness and diversity, and cannot hinge on how novelty is defined.

A larger corpus discriminates where ZINC does not. Recomputing novelty against ChEMBL (1,488,640 canonical molecules, roughly six times ZINC 250K) across all 150 model–task cells lowers it in every one: the median drop is 0.135, with 97 cells falling by more than 0.1, 31 by more than 0.3, and 7 by more than 0.5. The effect belongs to the corpus rather than to any one model—mean drops run from 0.084 (LLaMA3-8B) to 0.380 (GPT-5), and while the largest single drop is GPT-5 on DRD2 ($1.000 \rightarrow 0.287$), LLaMA3.1-70B on SA ($1.000 \rightarrow 0.398$) and DeepSeek-V3 on LogP= 5 ($1.000 \rightarrow 0.447$) are close behind. The two measures are uncorrelated ($\rho = -0.02$), which is the point rather than a contradiction: ZINC-based novelty spans only 0.965–1.000 and so has almost no ordering to preserve, whereas ChEMBL-based novelty spans 0.287–1.000, and 30 of the 134 cells scoring at least 0.99 against ZINC fall below 0.70 against ChEMBL. Binary novelty against a 250K corpus therefore carries little information in either direction—which is exactly why the preceding check matters, and why we report uniqueness and diversity separately rather than folding the dimension into one number. GPT-5 on DRD2 remains the sharpest case: entirely novel against ZINC but largely overlapping known ChEMBL molecules (Table 27).

Scope of the reference corpus. The corpus used for novelty is a fixed global one, not any model's actual training data. Novelty here means novelty relative to that corpus, and must not be read as evidence that a molecule was absent from a model's pretraining set.

H Alternative Diversity Definitions

Average pairwise Tanimoto distance is one of several defensible ways to measure molecular diversity, so we verify that our conclusions do not depend on that choice. We compute five complementary measures on the same generated sets for all 10 models $\times$ 15 evaluation tasks $\times$ 5 runs: Fréchet ChemNet Distance (FCD) and NSPDK MMD, which compare the generated distribution against the ZINC250K reference; unique scaffold ratio, a scaffold-level view of within-batch variety; #Circles-AUC (Xie et al., 2021), which measures chemical-space coverage; and Hamiltonian diversity (HamDiv) (Hu et al., 2024), a circuit-based measure. Table 7 reports per-model means.

Agreement with the reported diversity ranking. At the model–task level ($n = 150$), #Circles-AUC agrees closely with uniqueness ($\rho = 0.962$), internal diversity ($\rho = 0.932$), and the aggregate divergent creativity score ($\rho = 0.974$). #Circles-AUC and HamDiv are nearly interchangeable ($\rho = 0.999$), so HamDiv reproduces the same ordering. Across models, unique scaffold ratio agrees with uniqueness ($\rho = 0.94$) and diversity ($\rho = 0.89$), while FCD and NSPDK MMD are correspondingly negative ($\rho = -0.94$ and -0.96). The diversity conclusions in the main text therefore hold under distributional, scaffold-level, coverage-based, and circuit-based views alike.

Two measures that do *not* agree, and why that matters. Novelty runs weakly *against* chemical-space coverage ($\rho = -0.276$ against #Circles-AUC) where uniqueness and internal diversity run strongly with it ($\rho = 0.962$ and 0.932), reinforcing the point made in Appendix B.5 that novelty carries information the within-batch measures do not. Scaffold *similarity* also fails to track the

Table 7: Complementary diversity measures, averaged over all 15 tasks per model. FCD and NSPDK MMD are distances from the ZINC250K reference (lower = more distinct); the remaining columns are larger-is-better. DC is the divergent creativity score used throughout the paper.

Model	FCD ↓	NSPDK ↓	Scaff. Sim. ↓	Scaff. Div. ↑	#Circles ↑	HamDiv ↑
LLaMA3-8B	35.3	1.828	0.677	0.468	0.735	0.696
LLaMA3-70B	61.7	3.376	0.786	0.134	0.576	0.515
LLaMA3.1-8B	42.7	2.192	0.734	0.392	0.728	0.685
LLaMA3.1-70B	66.1	3.515	0.871	0.135	0.603	0.539
Qwen3-8B	48.8	2.448	0.462	0.246	0.647	0.596
Qwen3-32B	42.7	2.113	0.622	0.357	0.675	0.631
GPT-3.5	41.8	2.217	0.515	0.292	0.667	0.618
GPT-4o-mini	45.8	2.350	0.791	0.229	0.613	0.561
GPT-5	63.6	3.097	0.459	0.098	0.689	0.646
DeepSeek-V3	52.1	2.739	0.412	0.185	0.645	0.593

other measures ($\rho = 0.16$ against uniqueness, -0.05 against diversity); it is a distributional-overlap statistic rather than a within-batch spread statistic, and we report the unique scaffold *ratio* instead when a scaffold-level diversity reading is wanted. Finally, the raw and size-normalized variants of #Circles-AUC and HamDiv disagree with one another ($\rho \approx 0.27$–0.32 across variants), so the normalization convention must be stated whenever these metrics are quoted; the values above use the normalized form and the agreement statistics use the raw form, matching each metric's original definition.

Absolute scale relative to specialized models. LLM-generated distributions remain far from the reference distribution in absolute terms. The lowest FCD we observe for GPT-5 is 47.858, whereas specialized generative models reach 2.257–33.546, and the lowest NSPDK MMD is 1.815 against 0.0015–0.1770 for those models (Jo et al., 2023). High divergent creativity in our framework therefore reflects broad within-batch spread, not proximity to a curated drug-like distribution.

I Compatible Constraints

Supporting analyses for the task-side results of Section 5.

Adding compatible constraints can increase success rate rather than reduce it. Following (Zhou et al., 2025; Liu et al., 2025), we evaluate three two-constraint tasks: QED+SA, BBB+QED, and HIA+QED; detailed results are in Appendix S. Contrary to the intuition from natural language generation—where additional constraints typically reduce SR (Li et al., 2024a)—adding the second constraint does not uniformly reduce success. For LLaMA3-8B and LLaMA3.1-8B the SRs of all three multi-constraint tasks exceed that of the single QED constraint; across the full model set the effect varies by model and combination.

To diagnose this pattern, we measure QED satisfaction while prompting for a *different* property. Under the SA constraint alone, QED SR reaches 0.522 for LLaMA3-8B and 0.552 for LLaMA3.1-8B, both exceeding the 0.488 and 0.498 obtained under the QED constraint itself; BBB behaves similarly (0.508 and 0.616). This indicates that SA, BBB, and HIA guide generation toward regions of chemical space that are already enriched for drug-like molecules, consistent with shared structural determinants among molecular properties (Pereira et al., 2021).

J Numerical Constraints

Supporting analyses for the task-side results of Section 5, covering how closely generated distributions track a requested LogP value.

LLMs respond to numerical molecular constraints at the level of the output distribution rather than by exact matching. Using LogP as a case study, the mode of the generated distribution shifts in the direction of the specified target across all five settings (Spearman $\rho = 0.927$, $p < 10^{-6}$;

Pearson $r = 0.901$, $p < 10^{-6}$, computed between target LogP and the measured LogP of generated molecules), but the distributions remain broadly dispersed and deviate substantially from the requested value. Density and scatter visualizations are provided in Appendix J.1.

J.1 Additional Visualizations

This section supports Section J. Figure 6 shows the probability density of the measured LogP of generated molecules under each target, with Gaussian fits overlaid: the mode tracks the requested target while the spread around it remains wide. Figure 7 shows the per-molecule scatter for DeepSeek-V3 at the two extreme targets, and Figure 8 summarizes the relationship across all five targets. We show only the extremes here; the intermediate targets $(-1, 1, 3)$ interpolate between them without adding a distinct pattern.

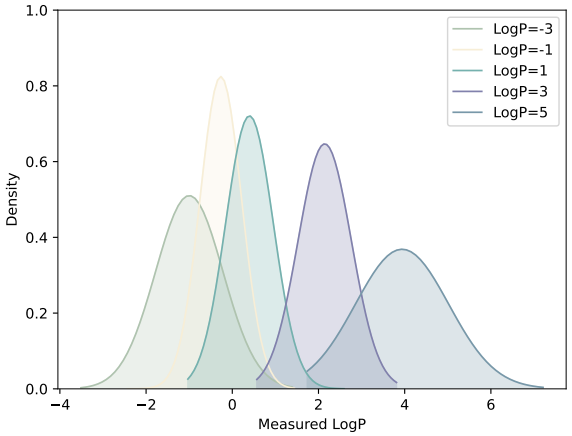
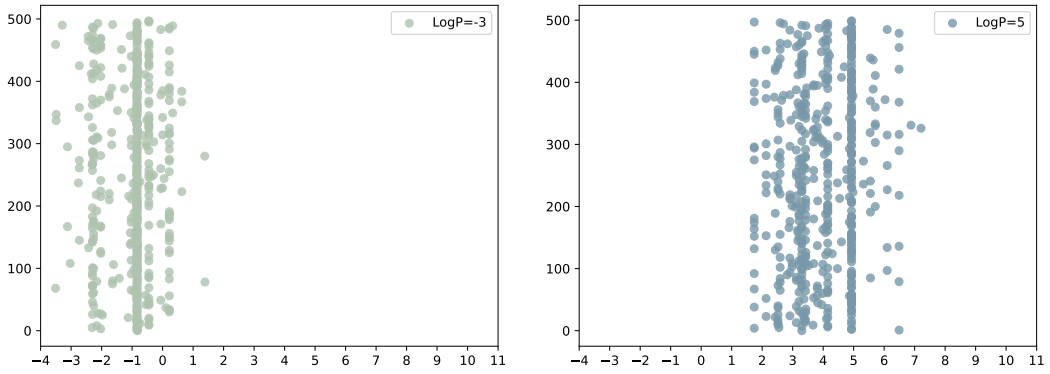


Figure 6: Generated LogP distributions shift consistently toward specified target values, but with substantial dispersion around each target.



(a) LogP=-3 (b) LogP=5

Figure 7: Scatter plots of molecular LogP distributions at the two extreme target conditions.

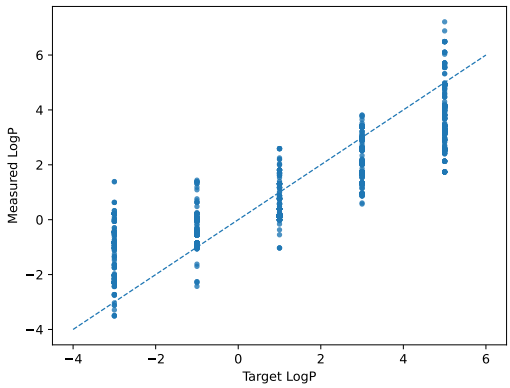


Figure 8: Conditional LogP distributions under discrete numerical constraints.

K Docking-Based Objectives

Every activity task in the main evaluation is scored by a computational oracle. To check whether the same qualitative profile appears under a structure-based score, we repeated the evaluation on two docking targets.

We dock GPT-3.5 generations with SMINA (Vinardo scoring, exhaustiveness 16) against two targets, 5-HT1B and the M2 muscarinic receptor (ACM2), using an existing benchmark's receptor structures, pocket coordinates and docking code, with 100 generations per target. A hit is a docking score at least as good as that benchmark's reported ZINC top-10% baseline (≤ -9.894 for 5-HT1B, ≤ -8.282 for ACM2).

Table 8: GPT-3.5 under docking-based scoring. The pattern matches the oracle-scored activity tasks: high validity and exploration alongside a low hit rate. Novelty-dependent quantities are omitted because this substudy predates the corpus revision of Appendix B; the columns shown do not depend on the reference corpus.

Target	Docking score	Hit rate	Validity	Uniqueness	Diversity
5-HT1B	-7.370 (1.349)	0.05	0.96	1.000	0.896
ACM2	-6.546 (0.893)	0.03	0.99	0.939	0.850

Convergent creativity is 0.219 and 0.172 respectively, against uniqueness and diversity that remain high. This is the same low-success/high-exploration profile the oracle-scored activity tasks produce in Section F, now observed on two structure-based docking targets.

L Additional Model-Side Analyses

Figure 9 gives the run-level view behind the family-balanced scaling endpoints in Figure 3b. This section supports Section 5.2. As shown in Figure 10, smaller models such as LLaMA3-8B and LLaMA3.1-8B exhibit validity ranging from approximately 0.64 to 0.96 across constraint settings, indicating unstable syntax-level generation; GPT-4o-mini shows similar variability (approximately 0.80 to 0.99). Larger models—LLaMA3-70B, LLaMA3.1-70B, GPT-3.5, GPT-5, and DeepSeek-V3—maintain near-saturated validity with minimal cross-task fluctuation. This pattern supports the model-side observation that some larger checkpoints occupy a more reliable syntax-level regime, while the comparison remains checkpoint-based rather than an isolated size intervention.

Size-varied checkpoint comparisons. The changes summarized in Figure 3B are paired by task family, with bootstrap confidence intervals and Benjamini–Hochberg correction across all reported comparisons (Appendix G). LLaMA3 8B→70B raises convergent creativity ($\Delta\mathrm{CC} = +0.191$, 95% CI $[+0.118, +0.268]$, $q = 0.004$) while lowering divergent creativity ($\Delta\mathrm{DC} = -0.406$, $[-0.492, -0.315]$, $q = 0.004$), a pattern that holds in all 11 task families; LLaMA3.1 behaves

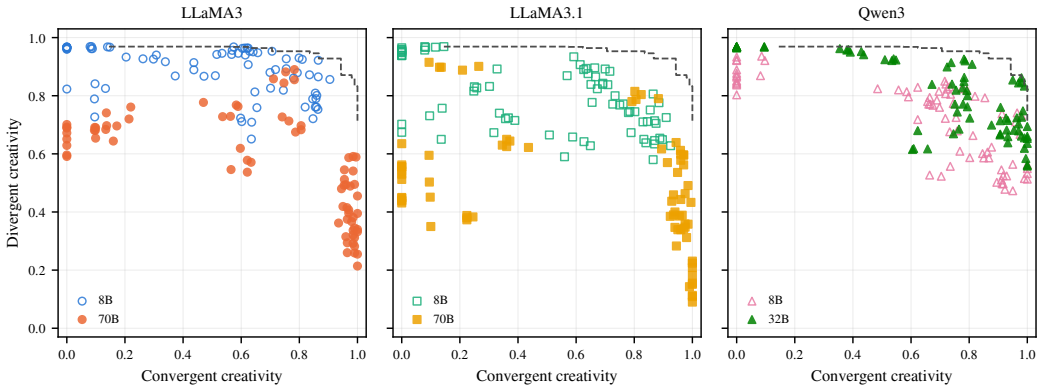


Figure 9: Run-level view of the scaling comparison summarized in Figure 3b. Each panel shows all runs from the two evaluated sizes of one model family against a descriptive pooled reference curve. LLaMA3 and LLaMA3.1 move right and down, while Qwen3 moves right without giving up divergent creativity. Unlike the main-text panel, which plots family-balanced endpoints, this view retains the spread across individual runs.

the same way (ΔCC $= +0.151$, $[+0.091, +0.215]$, $q = 0.004$; ΔDC $= -0.373$, $[-0.457, -0.280]$, $q = 0.004$) in 10 of 11. Qwen3 8B$\rightarrow$32B instead improves *both* dimensions in 8 of 11 families (ΔCC $= +0.073$, $[+0.023, +0.138]$, $q = 0.012$; ΔDC $= +0.068$, $[+0.028, +0.101]$, $q = 0.024$; ΔOverall Creativity, defined in Appendix B, $= +0.096$, $[+0.039, +0.176]$, $q = 0.016$). For the LLaMA families the loss of divergent creativity is not offset by joint gains: the Fully Creative metric, which requires a molecule to be simultaneously successful, novel, and unique, falls from 0.556 to 0.205 between LLaMA3-8B and LLaMA3-70B, and from 0.394 to 0.126 for LLaMA3.1.

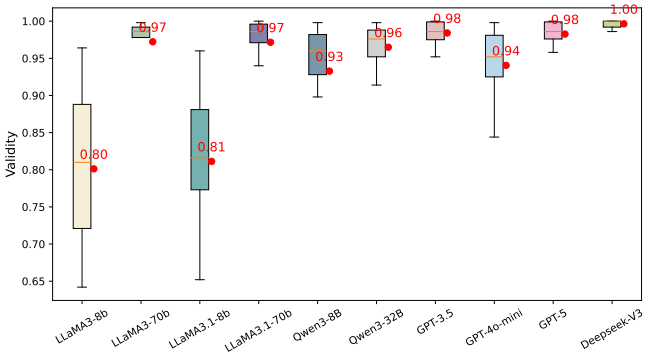


Figure 10: Larger models maintain near-saturated validity across diverse constraint settings, whereas smaller models exhibit substantial cross-task variability. The red line and dot denote the median and mean, respectively.

M In-Context Learning Results

This section provides additional evidence for the inference-time results of Section 5.3. Full per-metric results for the in-context learning comparison discussed in Section 5.3.

The example-selection contrast compares cluster medoids with uniformly random examples from the same active pool (GPT-3.5, 5 paired runs per task, identical seeds). Medoid selection raises success rate on GSK3β ($+0.342$) and DRD2 ($+0.116$) while lowering divergent creativity (-0.219 and -0.078); overall creativity is essentially unchanged in both ($+0.032$, $+0.006$). On JNK3 the direction reverses (ΔSR $= -0.124$, ΔDC $= +0.035$). The near rate exceeds the exact rate only marginally in every task, consistent with direct reproduction dominating broader analogue exploration

Table 9: ICL copy rate: the fraction of valid generated molecules reproducing a provided in-context example, exactly or at Tanimoto ≥ 0.8 (Morgan r=3, 2048 bits). Zero-shot runs are scored against the same example sets as a control and, as expected, never reproduce them. This diagnostic is defined only where in-context examples exist and is deliberately separate from novelty, which is measured against a fixed corpus.

Setting	Task	Exact copy rate	Near copy rate
Zero-shot	DRD2	0.000	0.000
	GSK3β	0.000	0.000
	JNK3	0.000	0.000
10-shot ICL	DRD2	0.366	0.373
	GSK3β	**0.655**	0.701
	JNK3	0.305	0.310
Medoid selection	DRD2	0.472	0.476
	GSK3β	0.647	0.678
	JNK3	0.437	0.447
Random selection	DRD2	0.318	0.329
	GSK3β	0.350	0.358
	JNK3	0.465	0.475

around the examples. Comparing the two selection strategies, medoid selection is copied more than random selection on DRD2 and GSK3β—the two tasks where it raises success rate—and not on JNK3, the one task where it does not.

Table 10: Effect of in-context learning on biological activity tasks. ICL substantially improves success rate but reduces structural exploration, yielding only marginal gains in Fully Creative. ZS: zero-shot; ICL: 10-shot in-context learning. A comparison of ICL performance with specialized generative models is provided in Appendix R.1.

Task	Validity	SR	Novelty	Uniqueness	Diversity	Fully Creative
DRD2 (ZS)	0.978 (0.015)	0.15 (0.019)	1.0 (0.0)	0.957 (0.018)	0.879 (0.006)	0.15 (0.019)
DRD2 (ICL)	0.9 (0.056)	0.542 (0.047)	1.0 (0.0)	0.688 (0.038)	0.83 (0.009)	0.26 (0.023)
JNK3 (ZS)	0.976 (0.013)	0.006 (0.005)	0.998 (0.005)	0.951 (0.008)	0.86 (0.005)	0.006 (0.005)
JNK3 (ICL)	0.732 (0.019)	0.316 (0.048)	1.0 (0.0)	0.763 (0.069)	0.848 (0.015)	0.142 (0.024)
GSK3β (ZS)	0.952 (0.016)	0.032 (0.019)	0.993 (0.006)	0.964 (0.012)	0.875 (0.004)	0.032 (0.019)
GSK3β (ICL)	0.84 (0.041)	0.698 (0.028)	0.994 (0.014)	0.407 (0.027)	0.643 (0.039)	0.2 (0.037)

N TEMPERATURE SENSITIVITY

This section supports Section 5.3. We sweep temperature in two independent model–task pairs, each with 5 runs of 100 samples per temperature.

LLaMA3-8B on SA (0.25–1.5; Figure 11). Divergent creativity increases with temperature: diversity grows near-monotonically across the full range, while uniqueness rises rapidly up to 1.25 and declines slightly at 1.5, suggesting saturation in structural variety. Convergent creativity is roughly constant from 0.25 to 1.0; above that both SR and validity decline. Overall creativity peaks at 1.25 (0.816).

GPT-3.5 on QED (0.2–1.2; Table 11). The same directional pattern appears—divergent creativity rises monotonically from 0.496 to 0.930 and convergent creativity falls—but the optimum sits lower, with 0.8 and 1.0 indistinguishable (0.835 and 0.836), and success rate peaks much earlier, at 0.5 (0.760), rather than holding flat to 1.0.

The same directional pattern appears in both tested model–task pairs, while the location of the optimum differs: a default near 1.0 is close to best in both, but the peak moves by roughly half a unit between them, so the value should be tuned rather than inherited.

Table 11: Temperature sweep for GPT-3.5 on QED. Overall creativity is indistinguishable at 0.8 and 1.0; success rate peaks at 0.5 and declines thereafter.

Temperature	Validity	SR	Uniqueness	Diversity	CC	DC
0.2	0.872	0.612	0.193	0.635	0.730	0.496
0.5	0.930	**0.760**	0.621	0.800	**0.841**	0.792
0.8	0.928	0.648	0.844	0.862	0.775	0.899
1.0	0.946	0.610	0.882	0.883	0.759	0.920
1.2	0.932	0.544	0.892	0.901	0.711	**0.930**

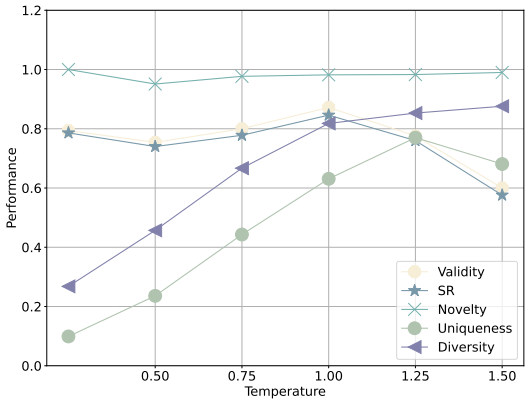


Figure 11: Temperature modulates the convergent–divergent trade-off for LLaMA3-8B on the SA task.

O Anchor-Conditioned Prompting

This section provides additional evidence for the inference-time results of Section 5.3, testing whether an explicit diversity-oriented prompt can raise exploration without losing constraint satisfaction.

Anchor-conditioned prompting improves exploration on some property tasks, but the net benefit is mixed. We compare three conditions for GPT-3.5 on QED (physicochemical), BBB (ADMET), and DRD2 (activity), each with 5 iterations of 100 samples. **C0** is the fixed prompt used throughout the paper. **C1** appends an explicit instruction to produce structurally diverse outputs. **C2** conditions on exemplars: each 100-sample run is split into five rounds, and rounds 2–5 are shown ten diverse anchor molecules selected by MaxMin picking *from that run's own earlier outputs*, with a request for dissimilar structures. C2 therefore introduces no external data and requires no retraining.

Table 12: Prompt interventions on GPT-3.5. C2 raises divergent creativity while holding convergent creativity roughly constant on QED and BBB, but both interventions reduce success rate on DRD2. FC is Fully Creative.

Task	**Condition**	**Validity**	**SR**	**Uniqueness**	**CC**	**DC**	**Creativity**	**FC**
QED	C0 (fixed prompt)	0.970	0.636	0.892	0.785	0.922	0.851	0.596
	C1 (diversity instr.)	0.942	0.698	0.796	0.811	0.867	0.838	0.550
	C2 (anchor-cond.)	0.936	0.682	**0.978**	0.799	**0.955**	**0.873**	**0.666**
BBB	C0	0.988	0.888	0.763	0.937	0.876	0.906	0.662
	C1	0.988	0.944	0.791	0.966	0.864	0.913	0.730
	C2	0.982	0.884	**0.961**	0.931	**0.951**	**0.941**	**0.848**
DRD2	C0	0.978	**0.150**	0.957	**0.382**	0.944	**0.601**	**0.150**
	C1	0.982	0.030	0.869	0.153	0.894	0.331	0.030
	C2	0.978	0.074	0.969	0.266	0.946	0.501	0.072

Table 12 shows that C2 improves overall creativity on QED ($+0.022$) and BBB ($+0.035$), reflecting large uniqueness gains (BBB: $0.763 \rightarrow 0.961$) while convergent creativity is essentially unchanged. The Fully Creative rate rises from 0.662 to 0.848 on BBB and from 0.596 to 0.666 on QED, but the same intervention is harmful where convergent creativity is already scarce: on DRD2 both C1 and C2 cut success rate (C1: $0.150 \rightarrow 0.030$; C2: $0.150 \rightarrow 0.074$), so overall creativity falls.

On a second, smaller model the divergent gain replicates but the net gain does not (Appendix P). C2 raises uniqueness on both QED ($0.819 \rightarrow 0.917$) and BBB ($0.851 \rightarrow 0.956$), but on BBB the 8B model loses success rate ($0.592 \rightarrow 0.456$) and overall creativity falls ($0.733 \rightarrow 0.659$). Anchor-conditioning therefore supports exploration, but it should not be read as a general solution to the trade-off.

P Cross-Model Validation of the Prompt Intervention

This section supports Section 5.3. To test whether the prompt-intervention result depends on the specific model, we repeated the C0/C1/C2 comparison on a local 8B checkpoint—an SFT-tuned LLaMA-3.1-8B—with 5 runs of 25 samples per task–condition. The sample size is deliberately small and C2 is sequential within a run, so later samples are not independent draws; these are directional observations and we make no significance claim.

Table 13: Prompt interventions on a local SFT-tuned LLaMA-3.1-8B. The uniqueness gain under C2 reproduces the GPT-3.5 result, but on BBB it is paid for in success rate and overall creativity falls. DRD2 sits at the success-rate floor and is uninformative.

Task	Condition	SR	Uniqueness	CC	DC	Creativity
QED	C0	0.248	0.819	0.321	0.874	0.520
	C1	0.216	0.879	0.293	0.901	0.497
	C2	0.240	**0.917**	0.307	**0.913**	0.525
BBB	C0	**0.592**	0.851	**0.600**	0.908	**0.733**
	C1	0.352	0.875	0.367	0.906	0.572
	C2	0.456	**0.956**	0.472	**0.947**	0.659
DRD2	C0	0.000	1.000	0.000	0.971	0.000
	C1	0.000	0.927	0.000	0.898	0.000
	C2	0.000	0.983	0.000	0.956	0.000

Of the 60 model–task–metric delta directions, 25 agree with the GPT-3.5 experiment. The component that reproduces is the one the intervention targets directly: C2 raises uniqueness in every task where uniqueness is not already at ceiling. The component that does not reproduce is the net effect on overall creativity, because on this model the diversity gain coincides with lower success rate. The available checkpoint is an SFT derivative rather than a vanilla base model, which may itself affect prompt sensitivity.

In-context example selection. A paired comparison contrasts cluster-medoid against uniformly random example selection for GPT-3.5 on the three activity tasks (5 paired runs, identical seeds). The resulting trade-off and copy-rate evidence are reported with the ICL results in Appendix M.

Serving route and its consequences. In the main benchmark, GPT-series models were accessed through the OpenAI API, while the other LLMs were served through OpenRouter. The GPT-3.5 experiments in this supplementary round were served through OpenRouter's `openai/gpt-3.5-turbo` route, whereas the GPT-3.5 results elsewhere in the paper were obtained directly from the OpenAI API. Two consequences follow. Within this round the route is constant across every compared condition—the example-selection contrast is paired run-for-run on identical seeds, and each temperature sweep is internally consistent—so the deltas we report are unaffected. Across rounds, however, the GPT-3.5/QED temperature sweep differs from the LLaMA3-8B/SA sweep in model, task, *and* serving route simultaneously, so the shift in the location of the creativity optimum should not be attributed to model or task alone. Separately, `gpt-3.5-turbo` is an unpinned legacy alias on both routes; we retain per-request metadata and system fingerprints, but alias drift remains a reproducibility limitation for any GPT-3.5 result in this paper.

This round comprises 6,625 generations (2,500 temperature, 3,000 example selection, 1,125 local intervention) from 5,508 API attempts, with all three experiments reported complete and all validation checks passing.

Q POST-TRAINING TRANSFER DIAGNOSTICS

This section decomposes and diagnoses the controlled post-training transfer result of Section 6; the analyses below are diagnostic unless stated otherwise. The SFT parent plus four LoRA adapters from one Llama-3.1-8B lineage—DPO, CrPO-CRE, CrPO-DIV and CrPO-QUA—are evaluated across the 15 canonical evaluation tasks, yielding 7,500 existing generations. No molecules were regenerated: the decomposition, descriptor and adapter analyses reuse stored per-molecule sidecars and the local adapter weights, and the likelihood analysis is forward passes only. The transfer table and CC decomposition use the task family as the paired unit ($n = 11$, or $n = 8$ excluding floor families), with exact sign-flip tests, 10,000-resample bootstrap intervals, and Benjamini–Hochberg correction within each analysis.

Table 14: Controlled comparison against the shared SFT parent, paired by task family ($n = 11$). Values are mean deltas relative to SFT. $\mathrm{SR} \mid \mathrm{valid}$ is constraint satisfaction among valid generations. Starred entries survive Benjamini–Hochberg correction at $q < 0.05$.

Adapter	Objective	ΔValidity	$\Delta(\mathrm{SR} \mid \mathrm{valid})$	ΔCC	ΔDC
DPO	Control	-0.089^{*}	-0.042	-0.075	$+0.019$
CrPO-CRE	Creativity/novelty/surprise	-0.048	-0.074	-0.056	$+0.018$
CrPO-DIV	Diversity	-0.117^{*}	-0.124	-0.129	-0.000
CrPO-QUA	Quality	-0.058	-0.046	-0.069	$+0.019$

Table 15: Exact additive decomposition of ΔCC under the family-balanced specification used in the main text, with and without the three activity families (DRD2, JNK3, GSK3β) whose convergent creativity sits at the floor for every variant. Excluding them raises the validity share for every variant. All eight contributions are negative, so validity is the larger contributor throughout without being the sole one.

Scope	Variant	n	ΔCC	validity	constraint	validity share
All 11 task families	DPO	11	-0.075	-0.058	-0.017	77%
	CRE	11	-0.056	-0.033	-0.023	59%
	DIV	11	-0.129	-0.086	-0.043	67%
	QUA	11	-0.069	-0.049	-0.020	71%
Excluding floors	DPO	8	-0.095	-0.078	-0.017	82%
	CRE	8	-0.069	-0.044	-0.025	64%
	DIV	8	-0.169	-0.118	-0.051	70%
	QUA	8	-0.087	-0.066	-0.021	76%

Generation-distribution shift. Median SMILES length rises from 36 under SFT to 44 (CRE, $q = 0.038$), 46.5 (QUA, $q = 0.006$), 49 (DPO, $q = 0.005$) and 54 (DIV, $q = 0.004$). Among the 3,481 valid molecules, median synthetic accessibility rises by $+0.85$ (DPO), $+0.95$ (CRE), $+1.23$ (DIV) and $+0.90$ (QUA), all $q \leq 0.021$ and on 13–14 of 15 tasks; median LogP and fraction Csp3 also rise. The shift is directional rather than dispersive—location effects dominate the ranking and only one scale effect reaches significance. Descriptors use the evaluator's own implementations, with agreement against the stored columns asserted rather than assumed.

Format compliance. The SFT parent almost never breaks the JSON output contract (compliance 0.922); every preference-tuned variant does (0.792–0.891). But conditioning on compliant responses closes only 12–46% of the CC gap, and validity remains down 0.047–0.091 among compliant responses, so format non-compliance is a real but secondary diagnostic signal. An early probe during this investigation suggested format was dominant; the fuller analysis does not support that reading.

Adapter geometry. With $\Delta W = 2.0\, BA$ over 64 modules, computed exactly through $r \times r$ algebra and verified against dense materialization, all six pairwise cosines fall in 0.093–0.106, against a cross-

layer null of 0.0006–0.0008 and a random rank-128 null of 0.0002–0.0003. The shared component is therefore real but small: it carries 0.98–1.11% of each adapter's update energy, and the residual against ΔW_{DPO} has norm 1.32–$1.35\times$ the adapter's own. Principal-angle overlap between the B column spaces is $2.6\times$ the random-subspace null. Same-seed pairs are indistinguishable from cross-seed pairs and the A-matrix cosine is ≈ 0 even within a seed, so the adapters do not share an initialization and the result is not a seed artifact. Update norms order DIV $>$ DPO $>$ QUA $>$ CRE, matching the damage order.

Likelihood. Teacher-forced negative log-likelihood of a fixed, seeded set of 100 reference molecules under three prompts, scored by each variant; nothing is sampled, so this is independent of decoding. Per-token NLL on the JSON scaffold rises from 0.209 under SFT to 0.745 (QUA), 0.857 (DPO), 0.999 (CRE) and 1.018 (DIV)—a factor of 3.6–4.9—while the SMILES body rises from 2.112 to 2.330–2.359, only 10–12%. Every variant raises NLL on 292–300 of 300 sequences ($q < 0.05$, two-sided sign test with bootstrap intervals, since an exact sign-flip is infeasible at this n). The likelihood ordering (CRE $\approx$ DIV $>$ DPO $>$ QUA) does not match the damage ordering (DIV $>$ DPO $>$ QUA $>$ CRE): CRE shows among the worst likelihood degradation but the least sampled damage. We report this as an unexplained non-result.

R Specialist Molecular Generators: External Context

These comparisons provide external context rather than matched evidence for the LLM analyses in the main paper: the systems below differ from the evaluated LLMs in training, objective, and sampling policy.

Specialist generators provide external context rather than evidence for the LLM-specific thesis. We re-evaluate three reference systems with our own metrics: GenMol (Lee et al., 2025), a task-optimized discrete-diffusion generator; REINVENT4 (Loeffler et al., 2024), a reinforcement-learning generator with an explicit diversity filter; and MolT5-base (Edwards et al., 2022), a molecule-finetuned language model prompted zero-shot. Because the sampling policy dominates these comparisons, we report two subsets for every method: **n100**, 100 molecules drawn at random from a run, and **top100**, the 100 highest-scoring molecules of that run.

Table 16: Success rates under matched n100 evaluation for the strongest specialist baseline and the strongest prompt-based LLM condition. Full matched results are in Table 17.

Method	DRD2	GSK3β	JNK3
GenMol n100	0.730	0.932	0.774
GPT-3.5 10-shot ICL	0.542	0.698	0.316

The compact comparison in Table 16 shows that task-optimized GenMol reaches substantially higher success rates than GPT-3.5 with 10-shot ICL on all three activity targets. MolT5-base, although trained on molecular data, fails entirely under the same zero-shot prompting protocol in the appendix table, so activity success depends more on the generation procedure than on molecular pretraining alone.

The appendix table also shows why the sampling subset must be stated. GenMol reaches saturated success under top100 selection, but only by selecting a less diverse subset. REINVENT4's diversity filter has little effect on random samples but preserves diversity among best-scoring candidates; disabling it lets success saturate while diversity and uniqueness collapse on top100 selection. These specialist comparisons are therefore baseline context, not evidence that LLM post-training or prompting has solved the LLM-specific convergent–divergent balance.

R.1 Published Numbers for a Wider Set of Specialist Methods

Table 17 above reports our matched re-evaluation. This subsection provides the complementary view: numbers as *reported in the original publications* for a wider set of specialist methods (GVAE + RL, GCPN (You et al., 2018), REINVENT (Olivecrona et al., 2017), RationaleRL (Jin et al., 2020b), f-RAG (Lee et al., 2024), and GenMol (Lee et al., 2025)), each obtained under that method's own generation and reporting protocol.

Table 17: Specialist generators and prompt-based LLMs under matched evaluation, on a random sample (n100) and the best-scoring sample (top100) of each run. "no DF" disables REINVENT4's diversity filter. Uniqueness is shown because it is where the filter's effect is clearest.

		DRD2			GSK3β			JNK3		
Method	**Subset**	SR	Div.	Uniq.	SR	Div.	Uniq.	SR	Div.	Uniq.
GenMol	n100	0.730	0.795	0.996	0.932	0.632	0.988	0.774	0.684	0.988
	top100	1.000	0.609	1.000	1.000	0.325	1.000	1.000	0.326	1.000
REINVENT4	n100	0.426	0.891	0.998	0.276	0.902	0.998	0.078	0.895	0.998
	top100	1.000	0.789	0.962	1.000	0.564	0.480	0.838	0.562	0.594
REINVENT4 (no DF)	n100	0.398	0.894	0.998	0.318	0.897	0.994	0.114	0.892	0.973
	top100	1.000	0.796	0.970	1.000	**0.285**	**0.176**	1.000	**0.300**	**0.192**
MolT5-base (zero-shot)	n100	0.000	0.772	0.532	0.000	0.810	0.616	0.000	0.859	0.750
GPT-3.5 (10-shot ICL)	n100	0.542	0.830	0.688	0.698	0.643	0.407	0.316	0.848	0.763

The two views should not be conflated, and the discrepancy between them is itself informative. Table 18 lists GenMol at DRD2 success 1.000 with diversity 0.818; our matched re-evaluation gives 0.730/0.795 on a random sample and 1.000/0.609 on the best-scoring sample (Table 17). Neither of our subsets reproduces the published pair, because the published number reflects a different selection policy over generated molecules. This is precisely why we state the sampling subset for every method rather than quoting published figures.

Read with that caveat, the literature numbers place prompt-based LLMs as comparable to or better than GVAE+RL and GCPN, while lagging behind REINVENT, RationaleRL, f-RAG, and GenMol. This is expected: specialist models are trained directly on full task-specific datasets and optimized for the target objectives, often with oracle-guided search, whereas the LLMs here operate in a general-purpose in-context learning setting without task-specific retraining.

Table 18: Success, novelty, and diversity on DRD2, GSK3β, and JNK3 tasks, *as reported in the original publications*. Each row follows its own method's generation and selection protocol, so these values are not directly comparable with the matched re-evaluation in Table 17.

Method	DRD2 Success	DRD2 Novelty	DRD2 Diversity	GSK3β Success	GSK3β Novelty	GSK3β Diversity	JNK3 Success	JNK3 Novelty	JNK3 Diversity
GVAE + RL	0.554	0.562	0.858	0.332	0.764	0.874	0.577	0.626	0.832
GCPN	0.401	0.128	0.880	0.424	0.116	0.904	0.323	0.044	0.884
REINVENT	0.981	0.288	0.798	0.993	0.610	0.733	0.985	0.316	0.729
RationaleRL	1.000	0.849	0.866	1.000	0.768	0.850	1.000	0.814	0.847
f-RAG	1.000	0.570	0.491	1.000	1.000	0.329	1.000	1.000	0.318
GenMol	1.000	0.997	0.818	1.000	0.956	0.843	1.000	0.983	0.861
GPT-3.5 + ICL	0.542	1.000	0.830	0.698	0.994	0.643	0.316	1.000	0.848

S DETAILED RESULTS

In this section, we present the detailed results of all models across all evaluation tasks in Tables 19 – 28. The reported values are the mean and standard deviation over five independent runs. Convergent and divergent creativity are omitted here because both are determined by the columns shown—convergent creativity is the geometric mean of validity and SR, and divergent creativity the geometric mean of novelty, uniqueness, and diversity. For the LogP tasks, we consider a generated molecule to be successful if its LogP differs from the specified value by at most 1.

Table 19: Detailed results for LLaMA3-8B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.82 (0.051)	0.774 (0.043)	0.997 (0.006)	0.882 (0.033)	0.884 (0.015)	0.856 (0.028)	0.676 (0.043)
HIA	0.828 (0.051)	0.78 (0.05)	0.981 (0.006)	0.776 (0.039)	0.833 (0.029)	0.831 (0.033)	0.586 (0.055)
QED	0.724 (0.074)	0.488 (0.044)	1.0 (0.0)	0.953 (0.007)	0.888 (0.007)	0.749 (0.037)	0.458 (0.044)
SA	0.872 (0.029)	0.846 (0.032)	0.985 (0.017)	0.631 (0.066)	0.818 (0.03)	0.828 (0.034)	0.518 (0.079)
DRD2	0.688 (0.041)	0.018 (0.011)	1.0 (0.0)	0.997 (0.006)	0.907 (0.004)	0.319 (0.049)	0.018 (0.011)
JNK3	0.642 (0.033)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.898 (0.006)	0.0 (0.0)	0.0 (0.0)
GSK3	0.642 (0.082)	0.002 (0.004)	1.0 (0.0)	0.994 (0.009)	0.903 (0.002)	0.058 (0.13)	0.002 (0.004)
BBB,QED	0.766 (0.038)	0.508 (0.058)	0.997 (0.006)	0.99 (0.016)	0.902 (0.007)	0.774 (0.028)	0.504 (0.054)
HIA,QED	0.718 (0.045)	0.492 (0.042)	0.991 (0.008)	0.967 (0.023)	0.9 (0.008)	0.752 (0.027)	0.466 (0.047)
QED,SA	0.81 (0.032)	0.568 (0.05)	0.994 (0.008)	0.783 (0.082)	0.829 (0.036)	0.765 (0.029)	0.394 (0.065)
LogP,-1	0.94 (0.016)	0.416 (0.03)	0.995 (0.01)	0.468 (0.064)	0.774 (0.025)	0.666 (0.024)	0.176 (0.048)
LogP,-3	0.924 (0.032)	0.01 (0.007)	0.997 (0.007)	0.629 (0.087)	0.823 (0.024)	0.231 (0.132)	0.008 (0.004)
LogP,1	0.964 (0.019)	0.758 (0.013)	0.98 (0.013)	0.542 (0.051)	0.837 (0.009)	0.807 (0.016)	0.34 (0.042)
LogP,3	0.904 (0.022)	0.222 (0.048)	0.989 (0.006)	0.81 (0.043)	0.855 (0.016)	0.626 (0.036)	0.214 (0.039)
LogP,5	0.776 (0.046)	0.106 (0.036)	0.997 (0.006)	0.918 (0.02)	0.875 (0.009)	0.51 (0.048)	0.106 (0.036)

Table 20: Detailed results for LLaMA3.1-8B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.888 (0.026)	0.87 (0.019)	0.996 (0.01)	0.538 (0.046)	0.74 (0.03)	0.803 (0.018)	0.458 (0.03)
HIA	0.816 (0.051)	0.794 (0.045)	0.996 (0.01)	0.59 (0.059)	0.733 (0.032)	0.779 (0.032)	0.458 (0.059)
QED	0.78 (0.037)	0.498 (0.052)	1.0 (0.0)	0.805 (0.07)	0.849 (0.031)	0.74 (0.022)	0.36 (0.052)
SA	0.874 (0.037)	0.858 (0.038)	1.0 (0.0)	0.383 (0.056)	0.69 (0.062)	0.744 (0.024)	0.318 (0.045)
DRD2	0.652 (0.043)	0.014 (0.011)	1.0 (0.0)	1.0 (0.0)	0.905 (0.004)	0.254 (0.147)	0.014 (0.011)
JNK3	0.654 (0.062)	0.0 (0.0)	1.0 (0.0)	0.964 (0.025)	0.885 (0.009)	0.0 (0.0)	0.0 (0.0)
GSK3	0.594 (0.065)	0.002 (0.004)	1.0 (0.0)	0.966 (0.018)	0.888 (0.008)	0.055 (0.123)	0.002 (0.004)
BBB,QED	0.766 (0.022)	0.574 (0.064)	0.997 (0.007)	0.82 (0.043)	0.846 (0.027)	0.765 (0.02)	0.434 (0.04)
HIA,QED	0.796 (0.022)	0.586 (0.036)	1.0 (0.0)	0.765 (0.059)	0.815 (0.021)	0.763 (0.018)	0.398 (0.065)
QED,SA	0.856 (0.028)	0.634 (0.076)	0.996 (0.008)	0.615 (0.033)	0.755 (0.018)	0.754 (0.024)	0.312 (0.058)
LogP,-1	0.96 (0.024)	0.36 (0.062)	1.0 (0.0)	0.347 (0.038)	0.746 (0.031)	0.61 (0.03)	0.132 (0.019)
LogP,-3	0.92 (0.025)	0.01 (0.01)	0.994 (0.013)	0.451 (0.061)	0.776 (0.03)	0.177 (0.163)	0.01 (0.01)
LogP,1	0.946 (0.033)	0.774 (0.059)	0.988 (0.016)	0.358 (0.017)	0.785 (0.021)	0.746 (0.015)	0.218 (0.026)
LogP,3	0.87 (0.025)	0.144 (0.019)	1.0 (0.0)	0.528 (0.039)	0.692 (0.024)	0.502 (0.025)	0.134 (0.021)
LogP,5	0.796 (0.031)	0.098 (0.027)	0.993 (0.015)	0.742 (0.06)	0.804 (0.022)	0.482 (0.039)	0.098 (0.027)

Table 21: Detailed results for LLaMA3-70B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.99 (0.007)	0.98 (0.012)	1.0 (0.0)	0.374 (0.049)	0.474 (0.045)	0.743 (0.028)	0.36 (0.043)
HIA	0.996 (0.009)	0.982 (0.025)	1.0 (0.0)	0.233 (0.045)	0.307 (0.075)	0.638 (0.046)	0.22 (0.037)
QED	0.988 (0.016)	0.946 (0.025)	1.0 (0.0)	0.178 (0.043)	0.223 (0.042)	0.572 (0.044)	0.136 (0.042)
SA	0.998 (0.004)	0.994 (0.009)	1.0 (0.0)	0.124 (0.034)	0.181 (0.045)	0.527 (0.046)	0.12 (0.032)
DRD2	0.924 (0.03)	0.024 (0.025)	1.0 (0.0)	0.564 (0.051)	0.655 (0.04)	0.224 (0.207)	0.022 (0.023)
JNK3	0.92 (0.016)	0.006 (0.009)	1.0 (0.0)	0.543 (0.026)	0.589 (0.024)	0.113 (0.156)	0.006 (0.009)
GSK3	0.88 (0.031)	0.012 (0.013)	1.0 (0.0)	0.49 (0.056)	0.54 (0.058)	0.176 (0.163)	0.01 (0.01)
BBB,QED	0.978 (0.008)	0.93 (0.01)	0.987 (0.017)	0.329 (0.026)	0.396 (0.027)	0.694 (0.018)	0.27 (0.037)
HIA,QED	0.99 (0.01)	0.94 (0.016)	1.0 (0.0)	0.21 (0.038)	0.273 (0.043)	0.608 (0.035)	0.158 (0.031)
QED,SA	0.996 (0.005)	0.97 (0.017)	1.0 (0.0)	0.163 (0.027)	0.177 (0.033)	0.547 (0.032)	0.136 (0.03)
LogP,-1	0.978 (0.022)	0.378 (0.028)	1.0 (0.0)	0.288 (0.039)	0.647 (0.047)	0.588 (0.022)	0.148 (0.025)
LogP,-3	0.984 (0.015)	0.012 (0.011)	1.0 (0.0)	0.405 (0.038)	0.745 (0.023)	0.224 (0.131)	0.012 (0.011)
LogP,1	0.994 (0.009)	0.614 (0.039)	0.973 (0.012)	0.457 (0.035)	0.769 (0.009)	0.739 (0.008)	0.276 (0.029)
LogP,3	0.982 (0.013)	0.582 (0.045)	0.995 (0.007)	0.831 (0.052)	0.785 (0.006)	0.809 (0.021)	0.492 (0.07)
LogP,5	0.986 (0.015)	0.306 (0.048)	1.0 (0.0)	0.781 (0.044)	0.548 (0.033)	0.642 (0.029)	0.204 (0.048)

Table 22: Detailed results for LLaMA3.1-70B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.992 (0.008)	0.97 (0.019)	1.0 (0.0)	0.24 (0.046)	0.328 (0.062)	0.646 (0.042)	0.216 (0.044)
HIA	0.998 (0.004)	0.998 (0.004)	1.0 (0.0)	0.068 (0.024)	0.081 (0.036)	0.413 (0.059)	0.068 (0.024)
QED	0.966 (0.011)	0.918 (0.019)	0.988 (0.028)	0.222 (0.037)	0.303 (0.049)	0.616 (0.033)	0.166 (0.033)
SA	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.058 (0.031)	0.061 (0.034)	0.38 (0.068)	0.058 (0.031)
DRD2	0.878 (0.029)	0.036 (0.029)	1.0 (0.0)	0.861 (0.025)	0.847 (0.004)	0.381 (0.078)	0.032 (0.022)
JNK3	0.94 (0.025)	0.0 (0.0)	0.98 (0.019)	0.318 (0.059)	0.394 (0.057)	0.0 (0.0)	0.0 (0.0)
GSK3	0.894 (0.025)	0.004 (0.005)	1.0 (0.0)	0.404 (0.058)	0.462 (0.054)	0.091 (0.125)	0.004 (0.005)
BBB,QED	0.976 (0.025)	0.894 (0.035)	0.974 (0.003)	0.402 (0.047)	0.53 (0.05)	0.743 (0.026)	0.302 (0.044)
HIA,QED	0.986 (0.015)	0.924 (0.027)	0.978 (0.031)	0.18 (0.046)	0.276 (0.064)	0.587 (0.048)	0.14 (0.036)
QED,SA	0.978 (0.016)	0.938 (0.026)	0.965 (0.032)	0.182 (0.018)	0.239 (0.031)	0.576 (0.019)	0.138 (0.026)
LogP,-1	1.0 (0.0)	0.044 (0.019)	1.0 (0.0)	0.144 (0.015)	0.366 (0.014)	0.274 (0.049)	0.032 (0.019)
LogP,-3	0.986 (0.011)	0.002 (0.004)	1.0 (0.0)	0.231 (0.048)	0.438 (0.042)	0.042 (0.094)	0.002 (0.004)
LogP,1	0.994 (0.005)	0.934 (0.011)	1.0 (0.0)	0.294 (0.027)	0.698 (0.015)	0.753 (0.012)	0.242 (0.016)
LogP,3	0.988 (0.013)	0.684 (0.059)	0.984 (0.001)	0.634 (0.03)	0.807 (0.005)	0.808 (0.019)	0.362 (0.054)
LogP,5	0.998 (0.004)	0.142 (0.028)	1.0 (0.0)	0.549 (0.02)	0.465 (0.014)	0.487 (0.021)	0.092 (0.016)

Table 23: Detailed results for Qwen3-8B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.988 (0.013)	0.93 (0.041)	0.974 (0.008)	0.47 (0.051)	0.809 (0.02)	0.829 (0.017)	0.398 (0.048)
HIA	0.976 (0.009)	0.89 (0.035)	0.992 (0.017)	0.248 (0.034)	0.564 (0.047)	0.694 (0.025)	0.196 (0.04)
QED	0.952 (0.019)	0.548 (0.029)	0.991 (0.014)	0.687 (0.022)	0.819 (0.013)	0.771 (0.011)	0.382 (0.026)
SA	0.998 (0.004)	0.998 (0.004)	1.0 (0.0)	0.228 (0.023)	0.669 (0.02)	0.73 (0.016)	0.228 (0.024)
DRD2	0.698 (0.185)	0.006 (0.005)	1.0 (0.0)	0.836 (0.09)	0.854 (0.016)	0.17 (0.156)	0.006 (0.005)
JNK3	0.806 (0.03)	0.0 (0.0)	1.0 (0.0)	0.84 (0.07)	0.879 (0.019)	0.0 (0.0)	0.0 (0.0)
GSK3	0.94 (0.027)	0.0 (0.0)	1.0 (0.0)	0.695 (0.042)	0.851 (0.021)	0.0 (0.0)	0.0 (0.0)
BBB,QED	0.96 (0.014)	0.338 (0.057)	0.988 (0.007)	0.667 (0.024)	0.827 (0.009)	0.68 (0.031)	0.252 (0.013)
HIA,QED	0.96 (0.012)	0.654 (0.082)	0.984 (0.012)	0.675 (0.042)	0.822 (0.009)	0.803 (0.016)	0.44 (0.022)
QED,SA	0.964 (0.019)	0.454 (0.038)	1.0 (0.0)	0.566 (0.054)	0.789 (0.006)	0.711 (0.025)	0.302 (0.057)
LogP,-1	0.916 (0.024)	0.624 (0.053)	1.0 (0.0)	0.423 (0.051)	0.789 (0.012)	0.723 (0.014)	0.266 (0.043)
LogP,-3	0.898 (0.029)	0.544 (0.047)	1.0 (0.0)	0.578 (0.034)	0.825 (0.017)	0.738 (0.02)	0.26 (0.037)
LogP,1	0.948 (0.02)	0.782 (0.036)	1.0 (0.0)	0.291 (0.02)	0.744 (0.02)	0.719 (0.013)	0.206 (0.015)
LogP,3	0.996 (0.005)	0.834 (0.018)	1.0 (0.0)	0.277 (0.032)	0.47 (0.025)	0.679 (0.016)	0.198 (0.026)
LogP,5	0.992 (0.008)	0.546 (0.078)	1.0 (0.0)	0.355 (0.039)	0.498 (0.052)	0.642 (0.043)	0.192 (0.048)

Table 24: Detailed results for Qwen3-32B

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.978 (0.011)	0.97 (0.02)	0.997 (0.006)	0.712 (0.026)	0.859 (0.005)	0.909 (0.007)	0.686 (0.026)
HIA	0.998 (0.004)	0.996 (0.005)	1.0 (0.0)	0.359 (0.053)	0.59 (0.039)	0.77 (0.026)	0.356 (0.053)
QED	0.964 (0.021)	0.628 (0.019)	1.0 (0.0)	0.685 (0.038)	0.852 (0.014)	0.806 (0.014)	0.476 (0.028)
SA	0.994 (0.009)	0.994 (0.009)	1.0 (0.0)	0.396 (0.033)	0.73 (0.011)	0.811 (0.014)	0.394 (0.034)
DRD2	0.944 (0.032)	0.16 (0.022)	1.0 (0.0)	0.972 (0.015)	0.898 (0.002)	0.608 (0.02)	0.134 (0.011)
JNK3	0.886 (0.046)	0.0 (0.0)	0.998 (0.005)	1.0 (0.0)	0.906 (0.001)	0.0 (0.0)	0.0 (0.0)
GSK3	0.914 (0.022)	0.004 (0.005)	1.0 (0.0)	1.0 (0.0)	0.908 (0.003)	0.121 (0.166)	0.004 (0.005)
BBB,QED	0.968 (0.016)	0.61 (0.05)	0.998 (0.005)	0.86 (0.022)	0.889 (0.008)	0.837 (0.022)	0.58 (0.038)
HIA,QED	0.976 (0.018)	0.656 (0.068)	1.0 (0.0)	0.51 (0.048)	0.741 (0.032)	0.76 (0.024)	0.298 (0.035)
QED,SA	0.982 (0.015)	0.57 (0.064)	0.997 (0.006)	0.628 (0.067)	0.834 (0.023)	0.775 (0.019)	0.352 (0.047)
LogP,-1	0.984 (0.011)	0.862 (0.018)	1.0 (0.0)	0.476 (0.046)	0.782 (0.017)	0.813 (0.013)	0.354 (0.034)
LogP,-3	0.94 (0.037)	0.282 (0.043)	1.0 (0.0)	0.892 (0.019)	0.895 (0.006)	0.69 (0.029)	0.274 (0.039)
LogP,1	0.994 (0.005)	0.954 (0.017)	1.0 (0.0)	0.463 (0.038)	0.794 (0.025)	0.835 (0.018)	0.426 (0.045)
LogP,3	0.992 (0.008)	0.866 (0.038)	1.0 (0.0)	0.383 (0.023)	0.759 (0.015)	0.783 (0.014)	0.306 (0.027)
LogP,5	0.96 (0.016)	0.44 (0.081)	0.996 (0.009)	0.446 (0.048)	0.592 (0.033)	0.643 (0.039)	0.2 (0.038)

Table 25: Detailed results for GPT-3.5

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.988 (0.008)	0.888 (0.016)	1.0 (0.0)	0.763 (0.014)	0.882 (0.01)	0.906 (0.005)	0.662 (0.011)
HIA	0.986 (0.017)	0.97 (0.014)	1.0 (0.0)	0.42 (0.027)	0.677 (0.026)	0.802 (0.01)	0.398 (0.029)
QED	0.97 (0.016)	0.636 (0.03)	0.997 (0.006)	0.892 (0.05)	0.882 (0.008)	0.851 (0.015)	0.596 (0.038)
SA	0.998 (0.004)	0.996 (0.009)	1.0 (0.0)	0.423 (0.033)	0.79 (0.015)	0.832 (0.011)	0.42 (0.028)
DRD2	0.978 (0.015)	0.15 (0.019)	1.0 (0.0)	0.957 (0.018)	0.879 (0.006)	0.601 (0.019)	0.15 (0.019)
JNK3	0.976 (0.013)	0.006 (0.005)	0.998 (0.005)	0.951 (0.008)	0.86 (0.005)	0.182 (0.166)	0.006 (0.005)
GSK3	0.952 (0.016)	0.032 (0.019)	0.993 (0.006)	0.964 (0.012)	0.875 (0.004)	0.393 (0.066)	0.032 (0.019)
BBB,QED	0.974 (0.011)	0.642 (0.036)	0.998 (0.005)	0.941 (0.024)	0.882 (0.009)	0.861 (0.01)	0.602 (0.048)
HIA,QED	0.978 (0.018)	0.416 (0.065)	1.0 (0.0)	0.671 (0.034)	0.836 (0.011)	0.724 (0.034)	0.392 (0.066)
QED,SA	0.964 (0.017)	0.704 (0.042)	0.995 (0.006)	0.929 (0.025)	0.875 (0.006)	0.876 (0.017)	0.674 (0.036)
LogP,-1	1.0 (0.0)	0.776 (0.013)	1.0 (0.0)	0.182 (0.029)	0.357 (0.039)	0.594 (0.026)	0.036 (0.033)
LogP,-3	1.0 (0.0)	0.006 (0.005)	1.0 (0.0)	0.366 (0.021)	0.672 (0.018)	0.151 (0.137)	0.006 (0.005)
LogP,1	1.0 (0.0)	0.992 (0.008)	1.0 (0.0)	0.086 (0.015)	0.189 (0.027)	0.501 (0.021)	0.078 (0.016)
LogP,3	1.0 (0.0)	0.672 (0.044)	0.996 (0.008)	0.458 (0.068)	0.709 (0.04)	0.749 (0.018)	0.212 (0.038)
LogP,5	0.998 (0.004)	0.314 (0.047)	1.0 (0.0)	0.798 (0.041)	0.76 (0.012)	0.687 (0.025)	0.23 (0.034)

Table 26: Detailed results for GPT-4o-mini

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.938 (0.028)	0.812 (0.052)	1.0 (0.0)	0.68 (0.052)	0.819 (0.017)	0.847 (0.017)	0.522 (0.033)
HIA	0.978 (0.008)	0.942 (0.011)	1.0 (0.0)	0.341 (0.046)	0.656 (0.029)	0.763 (0.018)	0.3 (0.038)
QED	0.912 (0.029)	0.552 (0.07)	1.0 (0.0)	0.656 (0.09)	0.801 (0.018)	0.755 (0.041)	0.394 (0.082)
SA	0.992 (0.008)	0.992 (0.008)	1.0 (0.0)	0.268 (0.029)	0.547 (0.041)	0.723 (0.022)	0.266 (0.03)
DRD2	0.888 (0.036)	0.0 (0.0)	1.0 (0.0)	0.796 (0.039)	0.814 (0.022)	0.0 (0.0)	0.0 (0.0)
JNK3	0.844 (0.034)	0.0 (0.0)	1.0 (0.0)	0.873 (0.047)	0.844 (0.014)	0.0 (0.0)	0.0 (0.0)
GSK3	0.796 (0.051)	0.0 (0.0)	1.0 (0.0)	0.903 (0.038)	0.849 (0.012)	0.0 (0.0)	0.0 (0.0)
BBB,QED	0.94 (0.023)	0.758 (0.013)	1.0 (0.0)	0.618 (0.053)	0.775 (0.016)	0.812 (0.02)	0.408 (0.055)
HIA,QED	0.97 (0.023)	0.62 (0.025)	1.0 (0.0)	0.484 (0.033)	0.765 (0.01)	0.746 (0.017)	0.324 (0.025)
QED,SA	0.95 (0.027)	0.606 (0.023)	1.0 (0.0)	0.562 (0.044)	0.764 (0.005)	0.756 (0.015)	0.4 (0.026)
LogP,-1	0.982 (0.015)	0.466 (0.027)	1.0 (0.0)	0.415 (0.022)	0.709 (0.008)	0.671 (0.013)	0.164 (0.031)
LogP,-3	0.988 (0.011)	0.014 (0.013)	1.0 (0.0)	0.579 (0.058)	0.748 (0.018)	0.203 (0.186)	0.014 (0.013)
LogP,1	0.998 (0.004)	0.57 (0.025)	1.0 (0.0)	0.597 (0.024)	0.773 (0.009)	0.763 (0.01)	0.318 (0.028)
LogP,3	0.98 (0.021)	0.442 (0.069)	1.0 (0.0)	0.725 (0.051)	0.792 (0.01)	0.738 (0.032)	0.33 (0.06)
LogP,5	0.952 (0.015)	0.246 (0.017)	1.0 (0.0)	0.862 (0.025)	0.828 (0.009)	0.657 (0.01)	0.224 (0.021)

Table 27: Detailed results for GPT-5

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.97 (0.019)	0.97 (0.019)	1.0 (0.0)	0.157 (0.012)	0.747 (0.038)	0.688 (0.017)	0.152 (0.013)
HIA	0.986 (0.011)	0.986 (0.011)	1.0 (0.0)	0.093 (0.021)	0.533 (0.078)	0.6 (0.03)	0.092 (0.02)
QED	0.96 (0.027)	0.732 (0.016)	1.0 (0.0)	0.29 (0.024)	0.752 (0.026)	0.71 (0.01)	0.248 (0.022)
SA	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.126 (0.025)	0.596 (0.025)	0.648 (0.023)	0.126 (0.025)
DRD2[†]	0.998 (0.004)	**0.998** (0.004)	1.0 (0.0)	**0.042** (0.008)	0.135 (0.029)	0.42 (0.027)	0.042 (0.008)
JNK3	0.918 (0.015)	0.002 (0.004)	1.0 (0.0)	0.44 (0.036)	0.458 (0.06)	0.049 (0.109)	0.002 (0.004)
GSK3	0.986 (0.017)	0.086 (0.015)	1.0 (0.0)	0.211 (0.038)	0.571 (0.055)	0.377 (0.017)	0.07 (0.014)
BBB,QED	0.986 (0.011)	0.966 (0.021)	1.0 (0.0)	0.177 (0.019)	0.639 (0.017)	0.686 (0.014)	0.164 (0.018)
HIA,QED	0.982 (0.015)	0.912 (0.031)	1.0 (0.0)	0.195 (0.028)	0.71 (0.017)	0.699 (0.023)	0.164 (0.03)
QED,SA	0.958 (0.013)	0.612 (0.056)	0.992 (0.017)	0.336 (0.043)	0.795 (0.022)	0.7 (0.031)	0.286 (0.043)
LogP,-1	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.03 (0.01)	0.333 (0.057)	0.461 (0.034)	0.03 (0.01)
LogP,-3	0.998 (0.004)	0.956 (0.019)	1.0 (0.0)	0.116 (0.017)	0.413 (0.024)	0.595 (0.019)	0.088 (0.016)
LogP,1	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.092 (0.011)	0.58 (0.026)	0.613 (0.015)	0.092 (0.011)
LogP,3	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)	0.036 (0.005)	0.203 (0.07)	0.438 (0.032)	0.036 (0.005)
LogP,5	0.998 (0.004)	0.08 (0.02)	1.0 (0.0)	0.11 (0.014)	0.437 (0.054)	0.318 (0.015)	0.028 (0.011)

[†] GPT-5 on DRD2 is an anomalous case with near-perfect SR but severe repetition (uniqueness $= 0.042$): roughly four distinct structures, each repeated about twenty-five times. Novelty does not register this, since none of these molecules appear in ZINC 250K.

Table 28: Detailed results for DeepSeek-V3

Task	Validity	SR	Novelty	Uniqueness	Diversity	Creativity	Fully Creative
BBB	0.986 (0.011)	0.982 (0.011)	1.0 (0.0)	0.205 (0.032)	0.695 (0.033)	0.716 (0.025)	0.198 (0.03)
HIA	1.0 (0.0)	0.996 (0.009)	1.0 (0.0)	0.098 (0.008)	0.258 (0.049)	0.54 (0.018)	0.094 (0.011)
QED	1.0 (0.0)	0.402 (0.045)	1.0 (0.0)	0.34 (0.047)	0.611 (0.045)	0.612 (0.037)	0.266 (0.027)
SA	1.0 (0.0)	0.998 (0.004)	1.0 (0.0)	0.1 (0.012)	0.523 (0.015)	0.611 (0.013)	0.098 (0.013)
DRD2	0.99 (0.01)	0.134 (0.03)	1.0 (0.0)	0.434 (0.042)	0.682 (0.033)	0.491 (0.031)	0.084 (0.019)
JNK3	0.992 (0.008)	0.004 (0.009)	0.998 (0.005)	0.929 (0.031)	0.868 (0.003)	0.073 (0.162)	0.004 (0.009)
GSK3	0.992 (0.008)	0.022 (0.018)	0.997 (0.006)	0.75 (0.035)	0.852 (0.007)	0.296 (0.171)	0.022 (0.018)
BBB,QED	0.99 (0.007)	0.398 (0.062)	1.0 (0.0)	0.303 (0.027)	0.718 (0.028)	0.613 (0.031)	0.218 (0.029)
HIA,QED	1.0 (0.0)	0.202 (0.029)	1.0 (0.0)	0.206 (0.026)	0.376 (0.028)	0.437 (0.025)	0.144 (0.023)
QED,SA	1.0 (0.0)	0.67 (0.033)	0.994 (0.013)	0.366 (0.021)	0.72 (0.02)	0.724 (0.013)	0.288 (0.03)
LogP,-1	1.0 (0.0)	0.886 (0.03)	1.0 (0.0)	0.158 (0.037)	0.611 (0.022)	0.655 (0.033)	0.1 (0.032)
LogP,-3	0.996 (0.005)	0.192 (0.052)	1.0 (0.0)	0.239 (0.042)	0.576 (0.031)	0.471 (0.032)	0.086 (0.019)
LogP,1	1.0 (0.0)	0.962 (0.011)	1.0 (0.0)	0.194 (0.023)	0.622 (0.024)	0.696 (0.018)	0.158 (0.024)
LogP,3	1.0 (0.0)	0.52 (0.06)	1.0 (0.0)	0.236 (0.038)	0.583 (0.032)	0.609 (0.033)	0.154 (0.035)
LogP,5	1.0 (0.0)	0.45 (0.069)	1.0 (0.0)	0.332 (0.059)	0.515 (0.04)	0.608 (0.021)	0.088 (0.024)