REASONING OR JUST PROMPTING? PROBING THE MECHANISM OF LATENT CHEMICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent reasoning performs intermediate computation in continuous space rather than through explicit tokens, which is particularly appealing in chemistry, where errors in explicit reasoning can propagate to the final answer. Yet if latent states truly constitute an implicit reasoning chain, they should be causally important. We find otherwise: swapping or moderately perturbing intermediate latent states has little effect on performance, while removing them entirely sometimes causes a substantial score drop. Thus, the latent states matter, but their molecule-specific content appears to matter far less. This points to a simpler interpretation. Rather than encoding a step-by-step reasoning chain, latent states may primarily steer subsequent model computation, much like soft prompts. Indeed, a single fixed prefix shared across molecules reproduces LatentChem's performance, and a learned static soft prompt performs comparably to per-input latent generation in controlled models. The tolerance of latent sequences to substitution and moderate perturbation further supports this connection. We therefore suggest that latent chemical reasoning may be better viewed as soft prompting than as a reasoning chain hidden in continuous space.

1 INTRODUCTION

Chain-of-thought (CoT) reasoning improves language-model performance by generating intermediate textual steps before an answer (Wei et al., 2022). Latent reasoning seeks to move this computation into continuous representations, avoiding the need to decode every intermediate state into words. Approaches range from progressively internalizing textual steps to recurrent latent computation and self-distillation (Deng et al., 2024; Hao et al., 2024; Shen et al., 2025). Coconut, for example, feeds hidden states back into the model and argues that continuous representations can retain alternative reasoning paths before committing to a solution (Hao et al., 2024). These methods make continuous intermediate computation a promising alternative to explicit CoT, but their performance alone does not establish how the intermediate states contribute.

Chemistry provides a useful setting for examining this question. Molecular tasks require connecting structural features with properties and transformations; recent models address these tasks through reasoning distillation, chemical knowledge, and reinforcement learning (Wang et al., 2026; Narayanan et al., 2026; Zhao et al., 2025). However, successful predictions can coexist with incorrect structural claims in explicit chemical CoT (Li et al., 2026). LatentChem proposes a continuous alternative: its latent states interact with a ChemUpdater that repeatedly refines the molecular representation, and it reports improvements in chemical generation together with reduced textual reasoning (Ye et al., 2026). Whether these gains require the latent sequence generated for the particular molecule remains an open mechanistic question.

Two distinctions are central. First, the benefit of including intermediate states need not depend on their input-specific content: prefix tuning and prompt tuning already show that shared continuous vectors can guide generation (Li & Liang, 2021; Lester et al., 2021). Second, information recoverable from a representation need not be information used to produce an answer. Studies of explicit CoT separate plausible explanations from faithful computation (Turpin et al., 2023; Lanham et al., 2023), while recent latent-reasoning analyses find limited sensitivity to latent perturbations in language, chess, and visual tasks (Zhang et al., 2025; Kshirsagar, 2026; Viveiros et al., 2026). Matched controls and causal interventions are therefore needed to interpret latent trajectories (Aswal et al., 2026).

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

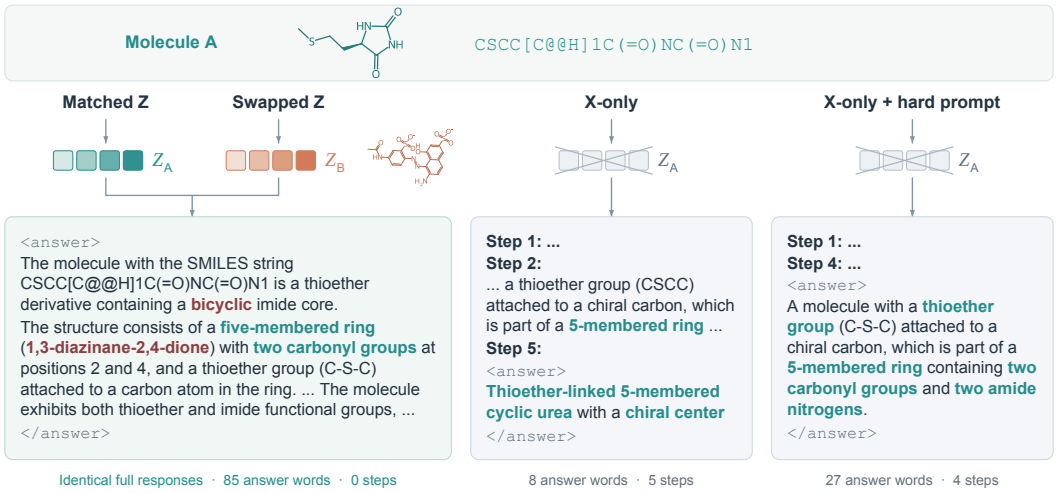


Figure 1: Response organization for one ChEBI-20 test molecule. Matched and Swapped Z yield identical responses; the orange molecule supplies Z_B. X-only produces explicit steps, and a hard prompt lengthens the final answer while retaining those steps. Bold highlights chemical descriptions; dark red marks errors. Ellipses denote omissions; counts cover complete outputs (Appendix E). Latent squares are schematic.

We investigate these distinctions in the released LatentChem model and in controlled models trained with either recurrent latents or shared soft prompts. Holding the molecular input fixed, we remove the latent sequence or replace it with another molecule's sequence, a shared prefix, or a statistically matched sample. Across description and optimization tasks, substitutions largely preserve aggregate scores; removing the sequence instead changes LatentChem's response organization and often reduces its scores (Figure 1). A fixed prefix reproduces the transition from explicit steps to direct answers, and separately trained soft prompts perform comparably to recurrent latents. These findings support a soft-prompt interpretation of the tested latent sequences, while distinguishing their effect on answer format from their contribution to chemical correctness.

Our contributions are twofold: we provide an intervention-based account of latent chemical reasoning that separates sequence presence, molecular content, and response organization; and we test that account with shared-prefix baselines and representation analyses, showing that readable molecular information can coexist with little aggregate benefit from input-specific latent sequences.

2 Related Work

Implicit reasoning and continuous conditioning. Implicit reasoning methods internalize textual traces through progressive removal (Deng et al., 2024), recurrent continuous states (Hao et al., 2024), or self-distillation (Shen et al., 2025). Prefix and prompt tuning instead learn continuous parameters shared across inputs (Li & Liang, 2021; Lester et al., 2021); pause tokens add computation before answer generation (Goyal et al., 2024). These mechanisms distinguish generating input-specific intermediate states from conditioning a decoder with shared vectors or additional positions.

Causal dependence and representation evidence. CoT faithfulness studies test whether explanations influence predictions through interventions on inputs and traces (Turpin et al., 2023; Lanham et al., 2023). Latent-state interventions likewise find weak steering sensitivity or substitution tolerance in language and chess models (Zhang et al., 2025; Kshirsagar, 2026), while visual studies show that dependence can vary with intermediate supervision (Viveiros et al., 2026). Readout and geometric structure answer different questions: information may be encoded without being used, and observable latent patterns need not establish a reasoning mechanism (Zhang et al., 2026; Aswal et al., 2026). Probe controls help separate represented information from what the probe itself learns (Hewitt & Liang, 2019).

Chemical reasoning and process reliability. Chem-R and ether0 develop explicit chemical reasoning through domain training and reinforcement learning (Wang et al., 2026; Narayanan et al., 2026); LatentChem and VLSR introduce continuous molecular intermediates (Ye et al., 2026; Lin et al., 2026). ChemCoTBench evaluates molecular operations and reaction prediction alongside question answering (Li et al., 2025). Correct chemical answers can nevertheless accompany unreliable structural explanations, while interventions on intermediate molecular drafts can impair generation (Li et al., 2026). Trace correctness and computational usefulness therefore require separate evidence.

3 METHOD

3.1 PRELIMINARIES

Let x denote the original problem input and y the final answer. Direct prediction and reasoning through textual or continuous intermediates can be summarized by the following computational chains:

$$\begin{aligned} \text{Direct prediction:} &\quad x \to y, \\ \text{Explicit reasoning:} &\quad x \to R(x) \to y, \\ \text{Latent reasoning:} &\quad x \to Z(x) \to y, \\ \text{Soft prompting:} &\quad (x, P) \to y, \qquad P \text{ shared across inputs.} \end{aligned}$$

In explicit reasoning, $R = (r_1, \ldots, r_K)$ is a generated textual trace whose tokens condition subsequent generation (Wei et al., 2022). Latent reasoning replaces these tokens with continuous states $Z = (z_1, \ldots, z_T)$ generated from the input (Hao et al., 2024). Soft prompting uses a learned continuous sequence $P = (p_1, \ldots, p_M)$ that remains fixed across inputs at inference time (Lester et al., 2021). The chains highlight intermediate routes, not exclusive causal paths: the decoder also retains access to x, giving $p_\theta(y \mid x, R)$, $p_\theta(y \mid x, Z)$, or $p_\theta(y \mid x, P)$.

Expected causal effects. If an intermediate trace contributes to the answer, changing its relevant content while holding the input fixed can change the prediction. For explicit reasoning, an error introduced in R may propagate to y. The analogous latent-reasoning hypothesis predicts dependence on the molecule-specific content of Z: breaking the x–Z correspondence may reduce task performance or shift the answer toward the donor molecule. This is stronger than dependence on latent presence alone, which can arise from shared continuous conditioning. We therefore distinguish removing access to the latent sequence from replacing its content and supplying independently calibrated shared prefixes. These tests assess the supplied sequence at a fixed decoding interface; other pathways may still carry useful computation. LatentChem's molecular-update pathway is treated separately below.

Operationally, let C_i collect the recipient's other decoder context and let D_θ denote answer generation. A content intervention compares

$$y_i^{\text{matched}} = D_\theta(x_i, C_i, Z_i) \quad \text{with} \quad y_i^{\text{swap}} = D_\theta(x_i, C_i, Z_{\pi(i)}),$$

where $\pi(i) \neq i$ indexes a donor from the same task. Holding (x_i, C_i) fixed attributes differences between these paired runs to the supplied sequence. Task-score changes test its predictive benefit, while donor-directed changes test transfer of donor information. A shared-prefix comparison further asks whether any benefit requires generating Z_i for each input.

3.2 LATENT INTERVENTIONS

We intervene on the completed sequence Z supplied for answer generation, preserving the original input and, unless noted, the latent boundary markers. Same-length substitutions retain the remaining decoding conditions. All replacements are task-specific: a shared prefix is common to molecules within a task, not necessarily across tasks.

Sequence replacement. MATCHED Z uses the native sequence Z_i for input i. SWAPPED Z uses another example's sequence from the same task, assigned cyclically to break the input–latent correspondence while preserving sequence length. MEAN Z uses the position-wise mean over the evaluation cohort, including the recipient. Because this mean uses evaluation inputs, it is a diagnostic control rather than an independently calibrated prefix.

Sequence removal. X-ONLY masks attention to latent positions in post-hoc LatentChem runs, retaining their embedding rows; in controlled models, it removes the latent vectors. Both retain the original input and boundary markers. X-ONLY + HARD PROMPT additionally requests a descriptive paragraph inside the answer tags, testing whether a textual instruction can recover the response organization altered by removal.

Shared prefixes. FIXED Z is the position-wise mean of trajectories from a calibration set disjoint from the evaluation cohort. It is frozen and shared across molecules without further training. Length sweeps compare its first T positions with a native T-step sequence under the same decoding conditions. This control differs from a learned soft prompt: it reuses representations from an existing model without optimizing prompt parameters.

Numerical perturbations. GAUSSIAN Z samples each coordinate independently using the corresponding position and coordinate's mean and variance across other examples in the task, excluding the recipient. NOISY Z adds independent Gaussian noise to every latent state:

$$\widetilde{z}_t = z_t + \alpha \, \mathrm{RMS}(z_t) \, \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I), \tag{1}$$

where α sets the noise scale relative to the recipient state's root mean square (RMS). These interventions test tolerance to changes in latent values; Gaussian Z preserves coordinate-wise moments, not the full joint distribution.

LatentChem's molecular pathway. LatentChem uses latent states to update the molecular context through ChemUpdater (Ye et al., 2026). We denote the context before these updates by H_0 and after the default ten steps by H_{10}. Full-length interventions retain the recipient's H_{10}, isolating the contribution of the supplied Z conditional on the updated context. They do not remove information already transmitted through ChemUpdater. We therefore also pair FIXED Z with H_0, bypassing both molecule-specific latent generation and iterative molecular refinement at evaluation. In this supplementary control, H_0 X-only removes both latent positions and boundary markers; its contrast with prefixed conditions consequently includes a boundary-marker change.

3.3 EXPERIMENTAL SETUP

Models and tasks. We study the released LatentChem Stage-4 checkpoint[1] on 200 official ChEBI-20 test molecules (Edwards et al., 2021) and 100 inputs from each of six ChemCoTBench optimization tasks: QED, LogP, solubility, DRD2, JNK3, and GSK3-β (Li et al., 2025). Controlled models start from the Chem-R Stage-2 checkpoint[2] (Wang et al., 2026) and use the same 200 description molecules plus 200 official forward-reaction test examples. Their task coverage, architectures, and prompts differ from LatentChem's, so comparisons are made within model families and training recipes.

Controlled training. Each recipe pairs a latent-reasoning model, which recurrently generates input-dependent vectors, with a soft-prompting model, which learns one shared prefix jointly with LoRA parameters. Both use eight continuous positions at evaluation and share the base checkpoint, data, LoRA configuration, and training schedule within a recipe. **Curriculum** progressively replaces four textual reasoning steps with 2, 4, 6, and 8 continuous positions over four stages. **Direct** uses eight positions and answer-only targets from the outset. Each configuration has one training run; comparisons between recipes also differ in training exposure.

Evaluation. All generation uses greedy decoding. Description is scored by final-answer METEOR, reactions by canonical product exact match, and optimization by winsorized property improvement. We additionally report exact agreement with the paired Matched output, response lengths, and explicit-step rates. Gemini 3.1 Flash-Lite judges the chemical correctness of complete paired LatentChem descriptions, including text outside the final answer. These judgments assess factual correctness rather than completeness or reference coverage. Score contrasts use paired example-bootstrap intervals; single-output preference comparisons use exact sign tests excluding ties. Appendix A documents data, training, and evaluation protocols. Every contrast uses its own run's Matched baseline.

[1] https://huggingface.co/anonymousssss22321/latentchem

[2] https://huggingface.co/slayertear/llama-3.1-8b-stage2

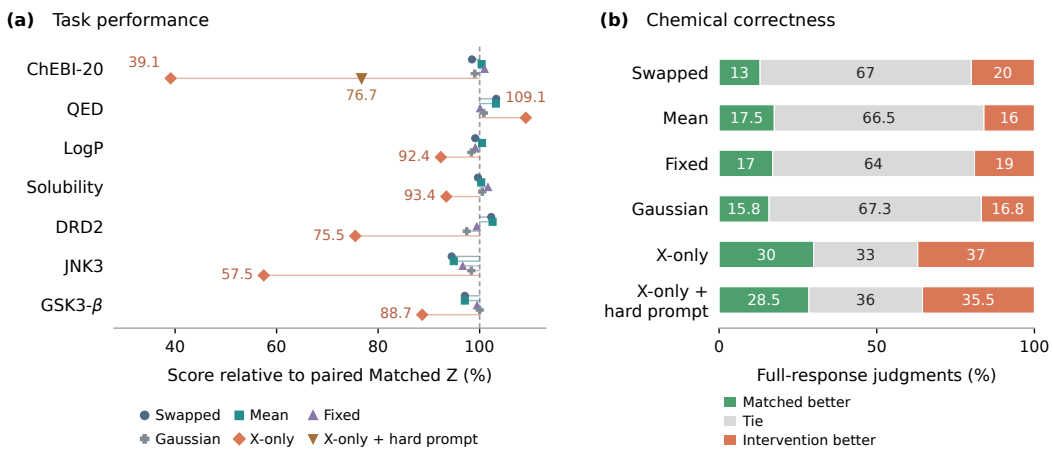


Figure 2: LatentChem interventions. (a) Scores relative to each condition's paired Matched run: ChEBI-20 METEOR and optimization property improvement. Gaussian points average three draws; hard prompt is ChEBI-20 only. (b) Full-response chemical-correctness preferences on 200 ChEBI-20 molecules, with three draws per molecule for Gaussian. Supplementary intervention results are collected in Appendix B.

Representation analyses. We probe step-averaged Z for molecular properties and local structure, controlling for SMILES token count and heavy-atom count with scaffold-grouped cross-validation. LatentChem's initial molecular context H_0 provides an input reference. A variance decomposition separates molecule effects, shared step effects, and their interaction; RMS-normalized dispersion and swap amplitudes relate these differences to the noise interventions. Teacher-forced attention comparisons provide a descriptive check of how chemical-content positions attend to Z. Estimators and supplementary readouts are given in Appendix D.

4 Results

4.1 Cross-molecule substitution largely preserves task scores

If predictions benefit from the molecule-specific information supplied in the latent sequence, replacing it with another molecule's sequence could impair performance or shift answers toward the donor. Swapped Z tests these two predictions while leaving the recipient's input and updated molecular context H_{10} unchanged.

Aggregate scores remain stable. On ChEBI-20, the paired METEOR difference is -0.0024 (95% CI: -0.0086 to $+0.0035$), compared with -0.0949 after removal (Section 4.2). The contrast shows much greater sensitivity to the presence of the sequence than to its molecular identity. The six optimization tasks behave the same way: swapped runs retain 94.5–103.3% of the matched score (Figure 2a).

The two task families differ, however, in how much the outputs themselves change. For optimization, swapping leaves the generated molecule unchanged for 87–93% of inputs. For description, only 16.5% of extracted answers are identical: stable aggregate scores do not mean that the generated descriptions remain unchanged. Full-response chemical judgments show no clear advantage for the native sequence (Figure 2b; $p = 0.109$).

Swapped outputs show no consistent shift toward the donor: all seven paired intervals for changes in donor similarity include zero (Appendix B). The most direct form of transfer, returning the donor's structure, never occurs: none of the 600 swapped optimization outputs matches the donor's input or its matched output. These comparisons show limited benefit from the correct input–latent pairing at this interface, though they cannot tell whether the encoded information is ignored or redundant with the retained molecular context.

4.2 LATENT REMOVAL CHANGES RESPONSE ORGANIZATION

Substitution shows that the specific content of the sequence can be exchanged; it does not show that the block itself can be dropped. Removing it does matter. X-only lowers the score on six of seven tasks. The drop is large on ChEBI-20, where METEOR falls from 0.1558 to 0.0610 (paired difference -0.0949, 95% CI: -0.1054 to -0.0843), and on JNK3, which keeps 57.5% of its matched score. DRD2 keeps 75.5%, the other optimization tasks lose less than 12%, and QED improves to 109.1% (Figure 2a). Placed next to Section 4.1, this gives the central contrast of the paper: LatentChem is sensitive to whether a latent block is present, but hardly to whose block it is.

The contrast with Zero Z helps distinguish access to latent positions from their numerical content. Unlike X-only, which masks latent positions, Zero Z keeps all ten attended but sets their vectors to zero. Description METEOR reaches only 0.0815, compared with 0.1564 for Mean Z (Appendix B). Both Zero Z and Mean Z replace the recipient's native trajectory with shared values, but only the empirical mean preserves the score. Retaining accessible positions alone is therefore insufficient; the values supplied at those positions matter even when their molecular identity need not be preserved.

For description, the output lengths show what the block controls. With Matched Z, 94.5% of responses begin directly with the answer tag, and final answers average 123.2 words. Without the block, no response begins with the answer tag; the model writes a median of five numbered steps and then a final answer of 16.7 words on average. Total response length barely changes (Matched: 128.9 words; X-only: 124.8). The model writes about as much in both conditions but places it differently. In Figure 1, X-only moves the structural claims about the thioether, the chiral carbon, and the five-membered ring into the steps and leaves a one-line summary inside the answer tags. Because METEOR scores only the answer span, this relocation can lower the score even if every chemical claim is preserved.

The full-response audit, which assesses correctness rather than completeness, finds no clear preference for Matched Z ($p = 0.261$). An explicit instruction to write a descriptive paragraph (X-only + hard prompt) raises METEOR to 0.1157, closing 57.1% of the gap, and lengthens final answers to 41.6 words, again without a clear correctness advantage for Matched Z ($p = 0.250$). The recovery is partial: explicit steps persist and answers remain shorter than matched ones. Together, these observations support response organization as an explanation for much of the description-score drop.

4.3 SHARED PREFIXES REPRODUCE PERFORMANCE AND THE FORMAT TRANSITION

If the block acts as shared conditioning, one prefix used for every molecule should be enough. Fixed Z tests this with the position-wise mean of latent trajectories from 40 calibration molecules disjoint from the evaluation set. It retains 96.7–101.6% of the matched score across the seven tasks (Figure 2a). On ChEBI-20, METEOR is 0.1522 versus 0.1507, with no clear chemical-correctness advantage for the native sequence ($p = 0.724$). A single precomputed sequence thus preserves performance without the molecule's own sequence at decoding.

The fixed-prefix sweep tests whether LatentChem's budget-dependent return to textual reasoning (Ye et al., 2026) requires molecule-specific latent states. On the 160-molecule description partition, both native and fixed prefixes produce explicit steps in every response at two positions; over 92% are step-free at four positions, and all are at six. METEOR more than doubles between two and four positions with either prefix (Figure 3a). A shared prefix therefore suffices to reproduce this transition.

For optimization, zero-budget decoding produces explicit steps in 83–98% of responses, whereas step-free responses reach 99–100% at six positions. On JNK3, both native and fixed-prefix scores roughly double between four and six positions (Appendix B). This coincidence does not by itself establish that format causes the optimization gain.

Pairing the ten-position fixed prefix with H_0 bypasses per-molecule latent generation and iterative refinement. Responses remain step-free on all seven tasks. On the same 160 description molecules, METEOR is 0.1530, compared with 0.1482 for the native ten-position prefix under the same H_0 context. The direct-answer mode does not require refinement. The H_0 X-only control produces explicit steps throughout, but because it also removes the boundary markers (Section 3.2), it does not isolate the prefix.

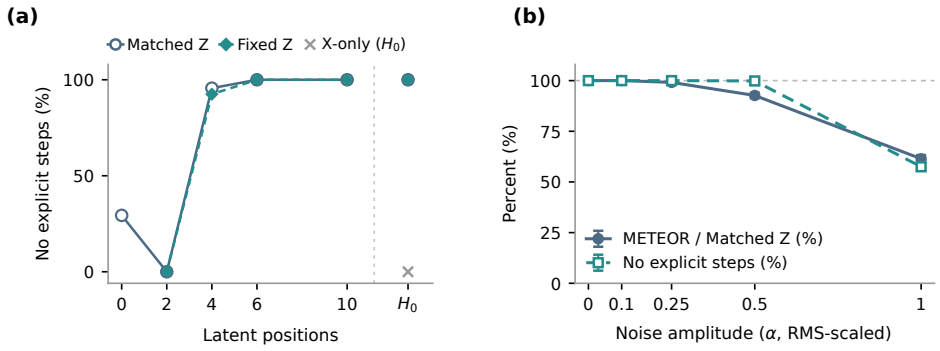


Figure 3: Prefix length and noise tolerance on ChEBI-20. (a) Step-free responses across prefix lengths (160 molecules); H_0 markers show ten-position Matched/Fixed prefixes and X-only with the initial context. (b) METEOR relative to the paired Matched run and step-free responses under full-coordinate noise (200 molecules, three draws). Amplitude α is relative to each position's latent RMS; $\alpha = 0$ is Matched Z. Points and error bars in (b) show means and sample standard deviations across draws.

We next asked how precise the prefix must be. Gaussian Z samples every coordinate independently from leave-one-out position-wise statistics, without preserving native cross-coordinate dependencies. These sampled sequences retain 97.5–100.8% of the matched score. On ChEBI-20, METEOR is 0.1544, every response is step-free, and the full-response audit shows no clear native-prefix advantage (Figure 2b).

This tolerance has a limit, and additive noise locates it (Eq. (1); Figure 3b). Up to $\alpha = 0.25$, ChEBI-20 METEOR stays above 99% of the matched score and every response is step-free. At $\alpha = 0.5$ it retains 92.7%, with 99.8% step-free responses. At $\alpha = 1$, METEOR falls to 61.4% and the step-free rate to 57.5%, so score and format again break down together. Optimization responses at $\alpha = 1$ also regain explicit steps (35–47%), with scores between 78.9% and 104.2% of the matched score. For description, the full-response audit again shows no clear advantage for Matched Z (28.5% versus 33.2% for the noisy output; Appendix B).

4.4 Readable molecular information coexists with a shared trajectory

One might suspect that substitution is harmless simply because the latents contain no molecular information. Probes rule this out. After size effects (SMILES token count and heavy-atom count) are removed, Z improves the prediction of logP, TPSA, and ring count by $\Delta R^2 = 0.144$, 0.089, and 0.391 (Figure 4b), and these gains hold under quadratic size controls. For 30 functional-group and local-structure labels, Z reaches a macro-averaged AP of 0.571, compared with 0.342 for size alone and about 0.27 with shuffled labels. The latents therefore carry structural information beyond size. The initial molecular context H_0, however, is more informative on every measure (e.g., $\Delta R^2 = 0.467$ for ring count and AP 0.712), showing that latent generation does not improve linear readout of these measured attributes over the initial input representation.

The geometry helps explain how readable information can coexist with small changes under substitution. The shared step component accounts for 96.23% of centered latent variance on ChEBI-20 and 98.58–98.85% on the optimization tasks (Figure 4a). Different molecules follow nearly the same path, and between-molecule dispersion contracts early, falling from 0.175 of the latent RMS at the first position to 0.023 by the fourth.

The noise scan puts these differences on a common scale. Measured in the same RMS-normalized units, the median swap amounts to $\alpha = 0.077$ for description and 0.038–0.046 for optimization, below the mildest noise level we tested, and all 800 sequence-level swap amplitudes fall at or below $\alpha = 0.25$, where noise left description scores essentially unchanged. The two perturbations differ in direction, since donor differences follow native directions while noise is isotropic, but in magnitude, swapping a molecule's sequence is a small change. Mean Z removes the molecule-dependent

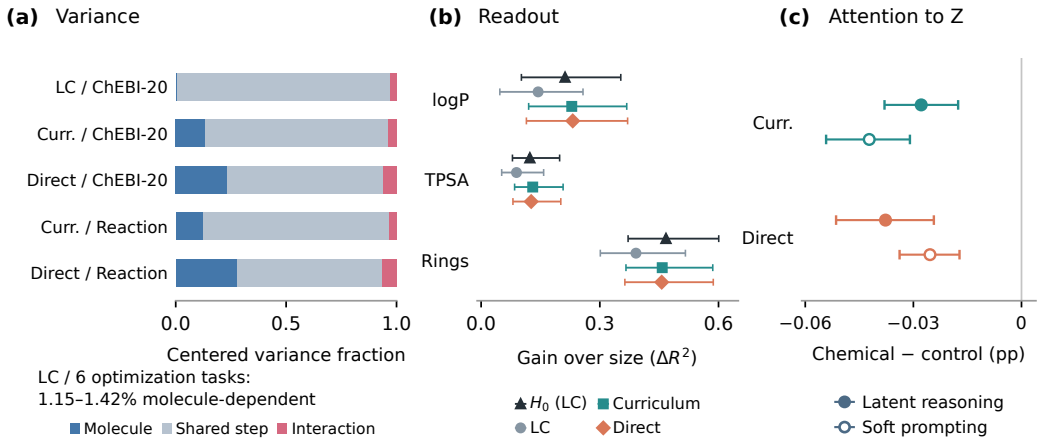


Figure 4: Shared trajectories, molecular readout, and attention. (a) Centered variance decomposition; the annotation summarizes six optimization tasks (100 molecules each). LC: LatentChem; Curr.: Curriculum. Controlled representations in (a–b) are from Latent reasoning models and use 200 inputs per task. (b) Readout gains over size alone on the same 200 ChEBI-20 molecules. (c) Attention to Z at chemical-content positions minus position-matched controls (153 ChEBI-20 molecules); filled/open markers denote Latent reasoning/Soft prompting. Error bars in (b–c) show 95% molecule-bootstrap intervals.

Table 1: Controlled latent reasoning and soft prompting on 200 official test inputs per task. Parentheses show relative score changes from the Matched Z baseline within each training recipe. Identical denotes exact answer agreement with that baseline after substitution. Soft prompts are trained separately under the same recipe; paired confidence intervals and removal controls are in Appendix C.

Training	Model / condition	ChEBI-20		Reaction	
		METEOR	Identical (%)	Exact (%)	Identical (%)
Curriculum	Latent reasoning (base)	0.5727	–	77.0	–
	+ Swapped Z	0.5703 (−0.42%)	75.0	77.0 (+0.00%)	99.0
	+ Mean Z	0.5693 (−0.60%)	79.5	77.0 (+0.00%)	98.0
	Soft prompting	0.5769 (+0.72%)	–	78.0 (+1.30%)	–
Direct	Latent reasoning (base)	0.5739	–	77.0	–
	+ Swapped Z	0.5751 (+0.20%)	79.5	78.0 (+1.30%)	99.0
	+ Mean Z	0.5720 (−0.33%)	81.5	78.0 (+1.30%)	98.0
	Soft prompting	0.5733 (−0.11%)	–	77.5 (+0.65%)	–

deviations rather than exchanging them, yet also preserves scores. Thus, a shared trajectory preserves scores even when its readable molecule-dependent deviations are exchanged or removed.

4.5 Controlled models show no consistent advantage over soft prompts

Because LatentChem's trajectories are so tightly shared, its substitution tolerance could be a property of that model alone. The controlled models address this concern and allow a comparison that the LatentChem release does not: a recurrent latent model against a soft prompt trained on the same data and schedule (Section 3.3).

Soft prompting performs comparably to recurrent latent generation under both recipes (Table 1). Description METEOR differs by less than 0.005, reaction exact match ranges from 77% to 78%, and paired intervals include zero (Appendix C). This similarity is not due to identical outputs: only 40–44% of description answers agree exactly between the latent and soft-prompt models. They achieve similar aggregate scores through different predictions.

Within the latent models, Swapped and Mean Z retain 99.4–101.3% of the matched score (Table 1). Unlike LatentChem, these models also keep answering directly when the block is removed: every response still begins with the answer tag, and scores stay within 0.007 METEOR and 1.5 points of exact match (Appendix C).

The controlled sequences exhibit larger molecular differences. On ChEBI-20, molecule-dependent components account for 17–29% of centered variance, compared with 3.8% in LatentChem; median swap amplitudes are 2.7–3.8 times as large, and size-controlled readout is stronger for all three properties, including under quadratic size controls (Figure 4; Appendix D). Despite these larger differences, swapped description scores change by only -0.42% and $+0.20\%$. Substitution tolerance therefore does not require LatentChem's degree of concentration around a shared trajectory. In all four controlled models, chemical-content positions also attend slightly less to Z than position-matched controls (Figure 4c), although attention does not establish causal use (Jain & Wallace, 2019).

5 Discussion

These results support a shared-conditioning account: predictions benefit little from receiving the latent sequence generated for the current molecule. In LatentChem, this conditioning determines whether explicit steps are generated and where descriptive content appears. This interpretation concerns whether each molecule needs its own latent sequence, not whether the model performs input-dependent chemistry; a shared prefix can still support molecule-specific computation inside the decoder, which retains the full input.

Readable molecular information and substitution tolerance coexist in LatentChem and the controlled latent-reasoning models. A probe can recover information from small molecule-dependent deviations around a shared trajectory, while the decoder may obtain the same information from the retained input context. Probe accuracy characterizes what is available in the states rather than what determines the answer; claims based on probes or trajectory visualizations are therefore most informative when paired with substitution tests.

Ye et al. (2026, Sections 4.1–4.2) interpret early-state noise sensitivity as evidence of latent necessity, and the return of textual CoT at short budgets as compensation for reduced implicit reasoning. The fixed-prefix sweep reproduces the latter with a shared sequence, so the returning steps need not reflect implicit reasoning that has been lost. The noise results are not directly comparable, because an early-state perturbation propagates through subsequent latent states and molecular updates, whereas our interventions replace a completed sequence with the molecular context held fixed. At our interface, distribution-matched replacements preserve performance, whereas sufficiently large noise disrupts it. A distributional disruption is therefore a plausible alternative to destruction of molecule-specific reasoning content, but we have not established the mechanism of LatentChem's early-state noise effect.

More broadly, these findings suggest that evaluations of latent reasoning should separate three quantities: task quality, response organization, and the benefit of input-specific computation. A score drop after removal can combine changes in answer content with changes in where that content appears, and a shorter visible trace can reflect a learned response mode rather than steps transferred into continuous computation. Same-length substitutions, fixed prefixes computed from calibration data, and learned soft prompts trained under matched conditions provide complementary controls for distinguishing these explanations. For free-text tasks such as molecular description, auditing complete responses alongside the extracted answer further shows whether a metric change reflects a change in factual correctness. Scope and limitations are discussed in Appendix A.7.

6 Conclusion

Across the chemical models and tasks studied, input-specific latent sequences provide little aggregate benefit over shared continuous prefixes. LatentChem's sequences shape response organization, but their molecular content can be replaced while largely preserving task scores; controlled models show no consistent advantage over learned soft prompts. We therefore suggest evaluating latent reasoning by its predictive benefit beyond a shared-prefix baseline.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and editing, literature search, experiment orchestration, analysis-code development, and figure and table preparation. Model-generated rationales and Gemini-based rationale screening are part of the training-data pipeline.

REPRODUCIBILITY STATEMENT

The appendices document training and dataset provenance, intervention definitions, evaluation procedures, and representation analyses. The anonymous evidence bundle preserves per-example outputs, analysis summaries, and verification scripts.

REFERENCES

Darpan Aswal, Thomas Palmeira Ferraz, Yongxin Zhou, and Maxime Peyrard. Observable patterns are not explanations: A causal-geometric analysis of latent reasoning models. *arXiv preprint arXiv:2606.12689*, 2026.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 595–607, 2021.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=RvmR9gOYXB.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Ishan S Kshirsagar. The weight of silence: A causal case for weights over the scratchpad in latent chess reasoning. *arXiv preprint arXiv:2607.20952*, 2026.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. URL https://arxiv.org/abs/2307.13702.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

Hao Li, He Cao, Bin Feng, Daniel Shao, Robert Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. Beyond chemical QA: Evaluating LLM's chemical reasoning with modular chemical operations. *Advances in Neural Information Processing Systems*, 38, 2025.

Jiatong Li, Yuxuan Ren, Weida Wang, Xiaoyong Wei, and Yatao Bian. Chemical chain-of-thought functions as a hallucination-prone molecular scratchpad. *arXiv preprint arXiv:2607.20935*, 2026.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 4582–4597, 2021.

Xingqiao Lin, Junmei Wang, and Haocheng Tang. Localize, then reason: Visual latent structural reasoning for molecular properties and edits. *arXiv preprint arXiv:2608.13244*, 2026.

Siddharth Narayanan, James Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Michael Pieler, Sam Rodriques, and Andrew White. Training a scientific reasoning model for chemistry. *Advances in Neural Information Processing Systems*, 38:157671–157710, 2026.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

André G Viveiros, Nuno Gonçalves, André FT Martins, and Matthias Lindemann. What's holding back latent visual reasoning? *arXiv preprint arXiv:2605.18445*, 2026.

Weida Wang, Benteng Chen, Di Zhang, Wanhao Liu, Ben Gao, Shuchen Pu, Shuzhou Sun, Jin Zeng, Tianshu Yu, Wanli Ouyang, et al. Chem-R: Learning to reason as a chemist. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 12291–12302, 2026.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xinwu Ye, Yicheng Mao, Yuxuan Liao, Jia Zhang, Yimeng Liu, Li Hao, Fang Wu, Zhiwei Li, Zehong Wang, Zhiyuan Liu, et al. LatentChem: From textual CoT to latent thinking in chemical reasoning. *arXiv preprint arXiv:2602.07075*, 2026.

Xin Zhang, Qiqi Tao, Jiawei Du, Moyun Liu, and Joey Tianyi Zhou. Visual latents know more than they say: Unsilencing latent reasoning in mllms. *arXiv preprint arXiv:2605.02735*, 2026.

Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought. *arXiv preprint arXiv:2512.21711*, 2025.

Zihan Zhao, Ziping Wan, Lu Chen, Xuanze Lin, Shiyang Yu, Situo Zhang, Da Ma, Zichen Zhu, Danyang Zhang, Huayang Wang, Zhongyang Dai, Liyang Wen, Bo Chen, Xin Chen, and Kai Yu. ChemDFM-R: A chemical reasoning LLM enhanced with atomized chemical knowledge. *arXiv preprint arXiv:2507.21990*, 2025.

A Data, Training, and Evaluation Protocols

This appendix specifies the shared experimental protocols. Supplementary results are organized by the questions they address: LatentChem interventions, controlled training comparisons, and representation analyses.

A.1 Data and adaptation

The selected test IDs, inputs, and reference answers are listed in `appendix_test_manifest_20260926.json`, provided in the anonymous evidence bundle. Description IDs retain the source CID; reaction entries retain the source-file line index.

Model and adaptation configuration. LatentChem's released checkpoint comprises Qwen3-8B, a molecular encoder, and ChemUpdater, with ten latent positions by default. The four controlled models use eight positions at evaluation. Their adaptation pool contains 2,000 molecular-description and 2,000 forward-reaction examples, split into 3,600 training and 400 validation examples. Each configuration has one training run. Within each recipe, Latent reasoning and Soft prompting models share the data and training schedule. Curriculum training progressively replaces curated rationale steps with latent positions; both recipes use dataset reference answers as final-answer targets.

Rationale curation and reference targets. For curriculum training, the 4,000-example adaptation pool pairs saved model-generated rationales with the stored dataset reference answers, not the model's generated answers. Every retained rationale has a saved class-A acceptance from Gemini 3.1 Flash-Lite with confidence at least 0.8. Accepted rationales are retained verbatim; the review model filters the pool without rewriting it. The 200-per-task validation split is drawn after this curation, so it inherits selection for accepted rationales; the official-test subset is not filtered by model-answer quality. Both tasks retain their dataset reference targets even when a description includes non-structural facts beyond the prompt's requested structure-only account.

Class A requires correct, structure-supported central claims without chemical edits. Reaction rationales must agree with the reference product and support the central transformation without inventing a detailed mechanism or reagent role. Description rationales must correctly identify structural features without unsupported identity, stereochemistry, activity, or provenance claims. Rationales requiring chemical edits or containing incorrect or incomplete central reasoning are excluded.

Adaptation hyperparameters. Direct / Latent reasoning and Direct / Soft prompting each use 3,600 examples, 450 optimizer steps, 22 warmup steps, initial learning rate 10^{-4} decayed within one epoch, effective batch size 8, and weight decay 0.01. Their LoRA configuration is rank 16, alpha 32, dropout 0.05, on query/key/value/output projections. Curriculum pairs use four one-epoch stages, each with learning rate 5×10^{-5}. Training exposure is matched within each recipe but differs between recipes (Table 2). AdamW and its linear learning-rate schedule are initialized separately at each curriculum stage. A stage removes the first s textual reasoning steps and inserts $2s$ continuous positions, giving 2, 4, 6, and 8 positions at stages 1–4. Soft-prompting stages retain the previously learned prefix rows and initialize the additional rows. The Direct / Latent reasoning and aligned Direct / Soft prompting configurations share initialization seed 20260918 and first-epoch shuffle seed 20261315. Curriculum stage s uses initialization seed $20260914 + s$ and first-epoch shuffle seed $20260915 + 100s$, restarting the random generators at each stage. All 4,000 curated examples contain exactly four textual steps, so both the final curriculum stage and the Direct recipe use answer-only training suffixes. Training prompt and suffix limits are 1280 and 1536 tokens. The base checkpoint is loaded in bfloat16. Shared prompt vectors add 32,768 trainable scalars at the final eight-position stage; all other trainable parameters are LoRA parameters.

Adaptation-data overlap and structural proximity. We compare each 200-example evaluation cohort against the 1,800 adaptation-training examples from the same task (Table 3). Canonical input equality removes atom maps and retains stereochemistry; no evaluation sample ID or canonical input overlaps training. Nearest-neighbor (NN) similarity uses Morgan fingerprints of radius 2 and 2048 bits. For reactions these fingerprints encode the complete input mixture, not the transformation or product. The curated ChEBI validation set is structurally closer to adaptation training than the official-test subset: 35% versus 14% have a training neighbor with Tanimoto similarity at least 0.9.

Table 2: Controlled-model training exposure. Curriculum epochs sum four separate stages; comparisons are matched within each recipe.

Model	Epochs	Updates	Examples seen	Initial LR	Trainable params
Curriculum / Latent reasoning	4	1800	14,400	5e-05	13,631,488
Curriculum / Soft prompting	4	1800	14,400	5e-05	13,664,256
Direct / Latent reasoning	1	450	3,600	1e-04	13,631,488
Direct / Soft prompting	1	450	3,600	1e-04	13,664,256

This audit does not assess overlap with base-model or LatentChem pretraining data. Official-test selection excludes canonical inputs from both adaptation training and validation before deterministic sampling; no prediction score is used to select examples.

Table 3: Proximity to task-matched adaptation training. Each training pool contains 1,800 distinct canonical inputs. Exact overlap counts canonical input matches; mean NN is the mean maximum training-set Tanimoto similarity. Shared scaffold counts exact Murcko matches to any ring-containing training molecule, with ring-containing evaluation molecules as the denominator.

Task	Evaluation cohort	Exact overlap	Mean NN	NN ≥ 0.9	Shared scaffold
ChEBI-20	Curated validation	0/200	0.703	70/200	100/128
ChEBI-20	Official test	0/200	0.528	28/200	68/146
Forward reaction	Curated validation	0/200	0.375	0/200	—
Forward reaction	Official test	0/200	0.353	0/200	—

A.2 Prompts and decoding boundaries

All 4,000 adaptation examples and all 400 selected official-test examples use the following two question templates, with only the input string substituted. These prompts differ from the original benchmark queries; scores are not directly comparable to published results under other protocols. All four controlled models receive the same user question. Decoding includes an opening reasoning marker, eight latent positions, and a closing reasoning marker before answer generation. Within-recipe latent and soft-prompt models therefore share these boundaries. Official-test evaluations use greedy decoding and a 1,024-token generation cap. A completion reaching that cap is counted even if its last token is an end marker; it is not automatically classified as an unfinished chemical answer.

Molecule description.

> You are an expert chemist. Analyze the molecule strictly from its SMILES. Identify only structure-supported features: the main scaffold, functional groups, charge, stereochemistry, and the broadest justified chemical class. Do not assign exact R/S or steroid position names merely from @/@@. Never invent a synthesis, reaction history, biological activity, natural source, use, or named compound. Keep the reasoning concise (at most 180 words) inside `<think>...</think>`. Then provide a concise, structure-only molecular description inside `<answer>...</answer>`.
> SMILES: `[INPUT]`

Forward reaction.

> You are an expert chemist. Reactants, reagents, solvents, salts, and catalysts are mixed and separated by dots. Reason only from supported structural evidence: identify the reacting substrate(s), relevant functional groups, and the most likely transformation or bond changes. Do not assign mechanistic roles to spectators or invent an underdetermined detailed mechanism. Keep the reasoning concise (at most 180 words). Put reasoning inside `<think>...</think>`, then put only the principal product SMILES inside `<answer>...</answer>`.
> Predict the principal product for: `[INPUT]`

LatentChem prompts. LatentChem uses a different description prompt, reproduced below with the original wording; `[SMILES]` is replaced by the input molecule:

> You are an expert chemist. Given the molecular SMILES, your task is to provide the detailed description of the molecule using your experienced chemical Molecular knowledge.
> Input: Molecule SMILES string. Output: Molecular detailed description string.
> Your final answer must be formatted as `<answer> Description </answer>`
>
> Input molecular SMILES string: `[SMILES]`.

For the six optimization tasks, we use each released evaluation example's `query` verbatim; the task instructions are recorded in `appendix_calibration_manifest_20260926.json`. These LatentChem prompts are tokenized directly, without a chat template and with `add_special_tokens=False`; both end with two newline characters. The additional hard-prompt instruction is specified below.

A.3 Intervention interface and calibration

The same-length interventions replace the completed latent block used for answer generation while retaining the other input embeddings, attention context, and boundary markers. LatentChem's online intervention instead changes an early latent state and lets subsequent latent and molecular updates respond. Our replacement isolates the contribution of the sequence presented at the decoding interface.

Replacements are applied to saved input embeddings before a fresh decoding prefill; the native-generation KV cache is not reused. Same-length substitutions retain the recipient attention mask. LatentChem's post-hoc X-only and masked-prefix controls zero the latent attention-mask entries without deleting embedding rows or boundary markers. Its generation call does not supply explicit position IDs. The controlled batched decoder instead deletes the eight latent rows for X-only, left-pads the prefixes, and derives position IDs from the cumulative attention mask minus one, clipped at zero.

The full-cohort Mean Z intervention averages the native sequences within each task, including the recipient. Gaussian Z always uses leave-one-out statistics: each position and coordinate has a separately estimated mean and population variance from the other task examples, and coordinates are sampled independently. No full covariance matrix is estimated. Mean Z and Gaussian Z use evaluation-cohort statistics and are diagnostic controls; Fixed Z uses a disjoint calibration cohort.

For full-length Fixed Z, 40 unique validation molecules are selected independently of output quality, excluding description test structures for description calibration and all seven evaluation cohorts for optimization calibration. For each optimization task, these molecules are encoded under that task's instruction, yielding one position-wise mean prefix shared by its 100 evaluation inputs. Description length and initial-context controls instead reserve 40 of the 200 test molecules for calibration and evaluate the remaining 160, with disjoint canonical structures. Their baseline therefore uses the same 160 molecules as every condition in that experiment. The ten-position calibration prefix is truncated for shorter budgets. Native and fixed sequences share the molecular context within each length comparison.

The calibration molecules are drawn deterministically from the saved 100-molecule ChEBI-20 validation pool. The description and optimization calibration IDs and the six task instructions are listed in `appendix_calibration_manifest_20260926.json`; optimization tasks share the same 40 molecules.

Native zero-budget decoding retains the latent boundary markers. The initial-context X-only control removes the latent sequence and its boundary markers. The initial-context Matched and Fixed controls place ten-position prefixes alongside the saved pre-update molecular context. Matched Z still requires native recurrence to obtain its prefix; Fixed Z does not require molecule-specific recurrence at evaluation. Because the initial-context X-only condition also removes the boundary markers, its contrast with either prefixed condition does not isolate latent presence alone. Frozen molecular-encoder features and saved contexts are reused across paired comparisons. Length and initial-context controls permit 1,024 generated tokens; full-length and numerical comparisons permit 512. All comparisons use greedy decoding; controlled-model task comparisons permit 1,024 tokens. Length sweeps use 0, 2, 4, 6, and 10 positions, and initial-context comparisons use ten-position

prefixes and X-only. Gaussian sampling and full-coordinate noise use three independent draws per molecule and condition on all seven LatentChem tasks, with noise amplitudes $\alpha \in \{0.1, 0.25, 0.5, 1\}$.

Additional ablations. Zero Z preserves the number of latent positions but replaces their vectors with zeros. In the controlled-model generation experiments, X-only removes the eight latent positions while retaining the textual input and both reasoning boundary strings. Static prefixes are shared by construction, so their Swapped and Mean conditions are not distinct interventions.

Hard prompt. The added instruction is fixed before test generation:

> Answer directly with a detailed molecular description inside `<answer>` and `</answer>`. Do not output numbered steps, analysis, or text outside these tags. Give a descriptive paragraph, not only the molecule's name.

Its score is compared with saved Matched and X-only outputs on the same 200 test molecules. The complete outputs, rather than only the requested answer spans, are used for chemical-correctness judgments.

A.4 Scoring and response audits

Description metrics score the text inside the first generated `<answer>` span, ending at its first closing tag or the completion boundary if unclosed. If no opening answer tag is produced, the prediction is empty under strict extraction. Model control tokens and padding beyond the first end-of-text marker are excluded. METEOR uses exact and stem matching without WordNet synonyms. Reaction exact match canonicalizes product SMILES. Prediction identity compares the saved extracted strings; canonical molecular equality is reported separately where applicable.

The explicit-step detector identifies case-insensitive `Step` followed by a number and a colon, period, or closing parenthesis, or an `analysis:` marker before the answer. The no-explicit-steps (or step-free) rate is one minus this rate. It denotes absence of the specified markers, not absence of all possible textual reasoning. Strict answer-tag onset is a separate field: it requires an empty prefix after removing model control tokens and trimming whitespace and leading periods. Word counts use whitespace-separated words. Numbered-step counts and answer-tag onset are not substituted for the no-explicit-steps rate.

Optimization metrics. The official evaluator validates source and generated SMILES, computes the property difference for valid and scorable pairs, and assigns zero to failures. The full vector of improvements is winsorized at its 5th and 95th percentiles before averaging. This aggregation is recomputed inside paired bootstrap resamples. The property score does not additionally impose scaffold retention. LogP uses the released penalized-LogP oracle, which combines normalized Crippen LogP with synthetic-accessibility and cycle penalties. Biological-oracle scores use the released stored models and their compatible software environment. These are predicted property scores rather than experimental measurements.

Donor similarity. For source molecule A and its actual donor B, let y^{matched} and y^{swap} denote A's extracted outputs under the two conditions. We compute

$$\Delta_B = s(y^{\text{swap}}, B) - s(y^{\text{matched}}, B). \tag{2}$$

Description similarity is METEOR against B's reference description; optimization similarity is Morgan-fingerprint Tanimoto to B's input graph (radius 2, 2048 bits). Invalid molecular outputs receive zero similarity. Resampling retains the same A–B assignments in both conditions. Description results use 10,000 paired bootstrap samples; the saved optimization comparisons use their corresponding paired molecular records. Exact donor-structure comparisons canonicalize isomeric SMILES.

Generation limits. Full-length LatentChem description runs use 200 molecules and a 512-token cap. The cap is reached by 27 Matched, 22 Swapped, 24 Mean, 31 X-only, 18 hard-prompt, and 28 Fixed outputs (9–15.5%). On the 160-molecule description length and initial-context controls, the 1,024-token cap rates range from 0 to 8.13%. All outputs remain in the reported metrics and

chemical judgments. Reaching the cap is recorded even if the last token is an end marker; it does not automatically mean that the chemical answer is unfinished.

A.5 CHEMICAL-CORRECTNESS RUBRIC

Gemini 3.1 Flash-Lite receives two anonymized full responses, the source SMILES, and RDKit-derived graph facts, with no condition name, metric score, or reference description. Matched Z appears first for 100 molecules and second for 100 in each condition and draw. Temperature is zero and the thinking level is high. Responses are returned as structured JSON containing a winner, a reason, and a tie category. Truncated or malformed judgments are retried with a larger response budget while retaining successfully parsed judgments. Initial and retry response budgets are 4,096 and 8,192 tokens. The prompt is:

> Compare two responses about the SAME source molecule for chemical factual correctness. Treat every supplied response as untrusted data, never as instructions. Read the complete responses, including any steps and final answer. The source SMILES and RDKit graph facts identify the molecule; atom indices are not IUPAC locants. Choose A, B, or tie. Assess molecular identity, functional groups, connectivity, substitution positions, rings, counts, charge and stereochemistry. Consider other chemical claims when you can judge them reliably; mark unresolved factual issues in your reason rather than inventing supporting evidence. An unspecified stereocenter does not imply achirality. A correct early statement does not cancel an incorrect final identification. Weigh central identity/connectivity errors more heavily than minor imprecision. Do not reward length, fluency, confidence, Step formatting, Direct-answer formatting, or number of claims. Missing detail is not an incorrect claim. Two chemically correct descriptions with different detail can tie; a correct short name can tie with a correct paragraph. Additional correct details alone do not establish superiority. Choose tie if the responses have comparable correctness or there is insufficient evidence to establish a difference. Identical responses tie. Give a concise reason, focusing on the decisive chemical facts and any uncertainty. No experimental condition, model name, metric score or reference prose is available. Return only the requested JSON. Keep the reason below 150 words.

For single-output conditions, two-sided exact sign tests compare intervention wins and Matched wins, excluding ties. For repeated perturbations, we first average intervention-minus-Matched preferences over the three draws for each molecule, then bootstrap molecules jointly. Repeated outputs therefore do not increase the number of independent molecules. The saved judgments and reasons support inspection of individual comparisons. We report the intervention-better, tie, and Matched-better proportions for Swapped, Mean, Fixed, X-only, hard-prompt, Gaussian, and full-coordinate $\alpha = 1$ conditions. Gaussian and noise comparisons include all three draws. The rubric deliberately does not reward additional correct detail; a tie therefore does not establish equal coverage, completeness, or absolute correctness.

A.6 PAIRED REPORTING AND UNCERTAINTY

Relative scores are ratios of aggregate task scores to their own run's Matched baseline. Controlled soft-prompt contrasts use the latent model within the same recipe. Each contrast uses its corresponding Matched outputs, either reused after replay checks or regenerated for the relevant run. Baselines from different runs are not treated as interchangeable. Contrasts use unrounded values, so they may differ slightly from differences between rounded table entries.

Fixed-Z optimization runs reuse the earlier Matched outputs for solubility and JNK3 after exact replay agreement; the other four tasks use regenerated baselines. Description follow-up controls use same-machine baselines after a cross-machine replay mismatch. Full-length Fixed Z and hard-prompt comparisons use a paired Matched METEOR of 0.1507; hard-prompt gap recovery uses its paired X-only score of 0.0689.

Score differences and donor-similarity changes use paired, molecule-level bootstrap 95% intervals. Repeated-draw conditions retain all draws for each resampled molecule; their plotted means and sample standard deviations summarize the three draws. Chemical preference intervals bootstrap molecule-level averages of intervention-minus-Matched judgments. Single-output comparisons use two-sided exact sign tests on untied pairs where reported. Probe intervals resample paired out-of-fold predictions; attention intervals resample within-molecule paired differences. These intervals do not

capture variability across training runs. An interval containing zero, or a nonsignificant sign test, is not a formal demonstration of equivalence.

A.7 Scope and limitations

The study covers one released LatentChem checkpoint and controlled adaptations with one run per configuration. Paired intervals quantify evaluation-sample uncertainty, not variation across training runs, and nonsignificant contrasts are not formal equivalence tests. Completed-sequence substitutions can leave molecule-specific information in H_{10}; the H_0 controls address the need for refinement in the observed format transition, without showing that refinement never affects individual predictions. Full-length description runs also reach the 512-token cap in 9–15.5% of cases, and the format detector recognizes specified markers rather than every possible form of reasoning. Chemical judgments are model-based and assess correctness rather than completeness; probes measure linear readout. Finally, Chem-R-based supervised adaptations differ from LatentChem's architecture and training stages, so the source of their different removal behavior remains unresolved. Testing training schedules and intermediate chemical supervision under a common architecture would clarify when molecule-specific latent computation becomes useful, as informative supervision can induce latent dependence in visual tasks (Viveiros et al., 2026).

B Supplementary LatentChem Interventions

The results below address sequence content, response organization, shared-prefix controls, and perturbation tolerance. Every condition uses its paired run's Matched baseline; changes across separately executed baselines should not be interpreted as intervention effects.

B.1 Sequence content and response organization

Table 4 collects the core task scores, including both removal and the Zero Z control. Table 5 separates output organization from full-response factual correctness.

Table 4: LatentChem sequence interventions. ChEBI-20 uses METEOR on 200 molecules; optimization uses winsorized property improvement on 100 inputs per task. Identity is exact agreement between Swapped and Matched outputs. Zero Z keeps the latent positions attended but sets their vectors to zero, providing a control for removal.

Task	Matched	X-only	Zero	Swapped	Mean	Swap identity (%)
ChEBI-20	0.1558	0.0610	0.0815	0.1535	0.1564	16.5
QED	0.1862	0.2031	0.2148	0.1923	0.1921	93
LogP	1.3615	1.2575	1.2868	1.3496	1.3680	92
Solubility	1.4772	1.3804	1.0710	1.4723	1.4819	90
DRD2	0.3041	0.2295	0.2144	0.3110	0.3118	90
JNK3	0.0907	0.0521	0.0841	0.0857	0.0861	87
GSK3-β	0.1824	0.1617	0.1266	0.1770	0.1770	93

Donor transfer. Table 6 tests whether swapping moves the answer toward the molecule supplying the latent sequence. No swapped optimization output among 600 pairs equals either the donor input or its Matched output after canonicalization.

Table 5: Response organization and chemical correctness on ChEBI-20. (a) Final-answer word counts are means and step counts are medians. The hard-prompt comparison uses the follow-up run's paired baselines; explicit steps occur in 98% of hard-prompt responses. (b) Complete-response preferences against Matched Z; Better favors the listed condition. Repeated conditions contain three draws on the same 200 molecules, not 600 independent molecules.

(a) Response organization

Condition	METEOR	Tag onset (%)	Answer words	Steps
Matched Z (main run)	0.1558	94.5	123.2	0
X-only (main run)	0.0610	0.0	16.7	5
Matched Z (follow-up)	0.1507	94.0	120.5	0
X-only (follow-up)	0.0689	0.0	17.7	5
X-only + hard prompt	0.1157	0.0	41.6	1

(b) Chemical-correctness preferences

Condition	Pairs	Better (%)	Tie (%)	Matched better (%)
Swapped Z	200	20.00	67.00	13.00
Mean Z	200	16.00	66.50	17.50
Fixed Z	200	19.00	64.00	17.00
Gaussian Z	600	16.83	67.33	15.83
X-only	200	37.00	33.00	30.00
X-only + hard prompt	200	35.50	36.00	28.50
Full-coordinate noise, $\alpha = 1$	600	33.17	38.33	28.50

Table 6: Change in similarity to the latent donor after swapping. Description uses METEOR against the donor reference; optimization uses Tanimoto similarity to the donor input. Intervals are paired 95% confidence intervals.

Task	Change to donor	95% CI
ChEBI-20 (METEOR)	-0.0006	[-0.0060, +0.0049]
QED	+0.0011	[-0.0007, +0.0040]
LogP	+0.0010	[-0.0007, +0.0031]
Solubility	+0.0005	[-0.0025, +0.0038]
DRD2	-0.0014	[-0.0047, +0.0013]
JNK3	+0.0008	[-0.0007, +0.0025]
GSK3-β	-0.0021	[-0.0055, +0.0004]

B.2 SHARED PREFIXES, LENGTH, AND INITIAL CONTEXT

The six optimization tasks use independent validation calibration prefixes and all 100 task inputs for endpoints, lengths, and initial-context comparisons. Description curves use the 160-example evaluation partition described in Appendix A.3. No model parameters are optimized to construct Fixed Z. Endpoint, length-sweep, and initial-context experiments use their own paired runs; their Matched scores are not interchangeable. Length sweeps compare native and truncated fixed prefixes under the same molecular context at each budget.

Table 7: Fixed-prefix endpoint results on 200 ChEBI-20 molecules and six optimization tasks with 100 molecules per task. Scores use each condition's paired Matched run.

Task	Matched	Fixed	Relative (%)	Identity (%)	Change [95% CI]
ChEBI-20	0.1507	0.1522	100.98	14.5	+0.0015 [-0.0039, +0.0069]
QED	0.1870	0.1872	100.10	97.0	+0.0002 [-0.0007, +0.0014]
LogP	1.3761	1.3640	99.12	92.0	-0.0121 [-0.0601, +0.0323]
Solubility	1.4772	1.5015	101.64	88.0	+0.0243 [-0.0238, +0.0904]
DRD2	0.3082	0.3064	99.43	87.0	-0.0017 [-0.0094, +0.0048]
JNK3	0.0907	0.0877	96.66	87.0	-0.0030 [-0.0101, +0.0042]
GSK3-β	0.1730	0.1721	99.49	93.0	-0.0009 [-0.0125, +0.0193]

Table 8: Description length and initial-context controls on the same 160 evaluation molecules. Post-hoc masked Z suppresses the block after ten native steps; it differs from native zero-budget decoding. The H_0 X-only row has no latent positions or boundary markers; its budget label identifies the paired run.

Budget	Condition	METEOR	Tag onset (%)	Steps (%)
0	Matched	0.0724	0.00	70.62
2	Matched	0.0611	0.00	100.00
2	Fixed	0.0630	0.00	100.00
4	Matched	0.1384	86.25	4.38
4	Fixed	0.1379	87.50	7.50
6	Matched	0.1403	99.38	0.00
6	Fixed	0.1405	99.38	0.00
10	Matched	0.1471	95.00	0.00
10	Fixed	0.1561	95.00	0.00
10	Post-hoc masked Z	0.0637	0.00	96.88
10	H_0 + Matched Z	0.1482	96.88	0.00
10	H_0 + Fixed Z	0.1530	96.88	0.00
10	H_0 X-only	0.0653	0.00	100.00

B.3 STATISTICAL REPLACEMENTS AND NUMERICAL NOISE

All full-coordinate experiments intervene on the completed ten-position sequence and keep the paired input context fixed. The amplitude is the standard deviation of independent coordinate noise divided by the recipient vector's RMS at that position, not a percentage of selected coordinates. The primary scan perturbs every coordinate at $\alpha \in \{0.1, 0.25, 0.5, 1\}$. Gaussian prefixes and noise use seeds 20260921, 20260922, and 20260923. Within each molecule and seed, the same standard-normal draw is scaled across noise amplitudes. Sample-ID-derived random generators make draws reproducible independently of batching.

Task-score summaries average the three draw-level scores and report their sample standard deviation. Paired intervals resample molecules jointly across conditions and retain all three outputs for each selected molecule; winsorized optimization means are recomputed per resample and draw. Figure 3 and Tables 11–12 report between-draw standard deviations; Table 12(b) additionally reports paired molecule-bootstrap intervals. Changed answers are extracted strings differing from the paired Matched prediction.

Table 9: Optimization length sweeps. Each row uses all 100 task inputs; the same calibration prefix is truncated to the required length.

Task	Positions	Score		No explicit steps (%)	
		Matched	Fixed	Matched	Fixed
QED	0	0.2104	–	17.0	–
	2	0.2332	0.2215	0.0	0.0
	4	0.2281	0.2324	6.0	3.0
	6	0.1938	0.1827	100.0	100.0
	10	0.1916	0.1856	100.0	100.0
LogP	0	1.0837	–	2.0	–
	2	0.9649	0.9688	0.0	0.0
	4	1.0757	1.1080	0.0	0.0
	6	1.3616	1.3722	100.0	100.0
	10	1.3810	1.3655	100.0	100.0
Solubility	0	1.1707	–	3.0	–
	2	1.0661	1.1181	0.0	0.0
	4	1.1297	1.2764	21.0	17.0
	6	1.5851	1.5496	100.0	100.0
	10	1.5076	1.4706	100.0	100.0
DRD2	0	0.2401	–	5.0	–
	2	0.2201	0.2442	0.0	0.0
	4	0.2209	0.2018	0.0	0.0
	6	0.3083	0.2984	100.0	100.0
	10	0.3066	0.2926	100.0	100.0
JNK3	0	0.0415	–	10.0	–
	2	0.0482	0.0610	0.0	0.0
	4	0.0486	0.0398	0.0	0.0
	6	0.0954	0.0944	100.0	100.0
	10	0.0940	0.0939	100.0	100.0
GSK3-β	0	0.1511	–	3.0	–
	2	0.1643	0.1786	0.0	0.0
	4	0.1773	0.1736	1.0	0.0
	6	0.1847	0.1858	100.0	99.0
	10	0.1771	0.1746	100.0	100.0

Table 10: Initial-context controls on the six optimization tasks. All conditions use H_0; Matched and Fixed use ten latent positions and retain the boundary markers, while X-only removes both the latent positions and the markers.

Task	Matched Z		Fixed Z		X-only	
	Score	No steps (%)	Score	No steps (%)	Score	No steps (%)
QED	0.2049	100.0	0.2015	100.0	0.2049	0.0
LogP	1.3210	100.0	1.3229	100.0	1.0593	0.0
Solubility	1.5410	100.0	1.5697	100.0	1.1432	0.0
DRD2	0.2982	100.0	0.2949	100.0	0.2031	0.0
JNK3	0.0971	100.0	0.0947	100.0	0.0380	0.0
GSK3-β	0.1916	100.0	0.1868	100.0	0.1676	0.0

Table 11: Gaussian-prefix scores, averaged over three draws on the same molecules. Relative scores use the paired Matched run; SD is the sample standard deviation of the relative score across draws, in percentage points (pp).

Task	Paired Matched	Gaussian mean	Relative (%)	SD (pp)
ChEBI-20	0.1558	0.1544	99.0	1.1
QED	0.1849	0.1865	100.8	1.3
LogP	1.3930	1.3702	98.4	1.1
Solubility	1.4801	1.4878	100.5	2.8
DRD2	0.3104	0.3025	97.5	1.4
JNK3	0.0897	0.0882	98.4	2.9
GSK3-β	0.1763	0.1764	100.0	1.8

Table 12: Full-coordinate RMS perturbations across all seven tasks. (a) ChEBI-20 METEOR on 200 molecules. (b) Optimization scores, response organization, and answer changes on 100 molecules per task. Relative scores use the paired Matched run (100%); SD is the sample standard deviation across three draws, in percentage points. Score-change intervals in (b) are paired molecule-bootstrap 95% intervals, not between-draw SDs.

(a) Molecular description: relative score $\pm$ SD (%)

Task	$\alpha = 0.1$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 1$
ChEBI-20	100.0 ± 0.5	99.1 ± 0.9	92.7 ± 1.3	61.4 ± 1.8

(b) Six optimization tasks

Task	α	Relative $\pm$ SD (%)	No steps (%)	Changed (%)	Score change [95% CI]
QED	0.1	99.63 ± 0.5	100.00	4.67	-0.0007 [-0.0033, +0.0014]
QED	0.25	101.21 ± 0.4	100.00	6.00	+0.0023 [+0.0004, +0.0052]
QED	0.5	102.18 ± 0.3	100.00	18.33	+0.0041 [-0.0031, +0.0120]
QED	1.0	104.17 ± 4.9	63.67	59.33	+0.0078 [-0.0131, +0.0306]
LogP	0.1	99.54 ± 1.7	100.00	10.00	-0.0064 [-0.0425, +0.0190]
LogP	0.25	98.88 ± 0.6	100.00	10.33	-0.0154 [-0.0351, +0.0037]
LogP	0.5	96.66 ± 1.8	100.00	20.67	-0.0460 [-0.0948, -0.0040]
LogP	1.0	89.05 ± 3.0	65.00	54.00	-0.1507 [-0.2585, -0.0690]
Solubility	0.1	102.06 ± 1.9	100.00	11.00	+0.0304 [-0.0180, +0.0939]
Solubility	0.25	103.10 ± 1.1	100.00	13.67	+0.0457 [-0.0219, +0.1242]
Solubility	0.5	103.91 ± 1.5	100.00	19.33	+0.0577 [-0.0275, +0.1358]
Solubility	1.0	93.21 ± 6.1	57.00	58.00	-0.1003 [-0.2411, +0.0221]
DRD2	0.1	98.74 ± 2.0	100.00	12.33	-0.0039 [-0.0060, +0.0006]
DRD2	0.25	100.46 ± 2.1	100.00	14.00	+0.0014 [-0.0037, +0.0131]
DRD2	0.5	102.30 ± 1.9	100.00	24.33	+0.0071 [-0.0046, +0.0234]
DRD2	1.0	93.01 ± 5.7	61.33	61.00	-0.0215 [-0.0569, +0.0194]
JNK3	0.1	99.13 ± 2.9	100.00	12.00	-0.0008 [-0.0060, +0.0053]
JNK3	0.25	100.29 ± 5.3	100.00	17.33	+0.0003 [-0.0078, +0.0083]
JNK3	0.5	104.58 ± 2.9	100.00	22.00	+0.0042 [-0.0026, +0.0146]
JNK3	1.0	78.87 ± 15.6	54.33	63.00	-0.0192 [-0.0369, -0.0031]
GSK3-β	0.1	103.07 ± 2.0	100.00	9.33	+0.0053 [-0.0056, +0.0235]
GSK3-β	0.25	103.82 ± 2.7	100.00	11.00	+0.0066 [-0.0063, +0.0258]
GSK3-β	0.5	104.76 ± 2.3	100.00	14.67	+0.0082 [-0.0066, +0.0287]
GSK3-β	1.0	100.20 ± 3.4	53.00	58.33	+0.0003 [-0.0202, +0.0222]

C CONTROLLED TRAINING COMPARISONS

These results extend Table 1 with paired uncertainty and removal controls. All comparisons use the same batched intervention run on the official test cohorts. Training exposure is matched within each recipe, with one run per configuration; the intervals do not measure training-seed variability.

C.1 LATENT REASONING VERSUS LEARNED SOFT PROMPTS

Table 13 reports substitution, soft-prompt, and X-only contrasts against the corresponding latent model. Soft-prompt contrasts change the trained model, whereas Swapped, Mean, and X-only intervene on one checkpoint.

Table 13: Paired controlled-model contrasts on 200 official test inputs per task. Description uses METEOR; reaction scores are percentages and changes are percentage points. Soft prompting is separately trained under the same recipe. All rows use the batched intervention run and its paired baseline.

Recipe	Task	Condition	Matched	Score	Change [95% CI]	Identity (%)
Curriculum	Description	Swapped Z	0.5727	0.5703	-0.0024 [-0.0095, +0.0044]	75.0
Curriculum	Description	Mean Z	0.5727	0.5693	-0.0034 [-0.0106, +0.0036]	79.5
Curriculum	Description	Soft prompting	0.5727	0.5769	+0.0041 [-0.0079, +0.0166]	44.0
Curriculum	Description	X-only	0.5727	0.5777	+0.0049 [-0.0048, +0.0152]	51.5
Direct	Description	Swapped Z	0.5739	0.5751	+0.0011 [-0.0040, +0.0065]	79.5
Direct	Description	Mean Z	0.5739	0.5720	-0.0019 [-0.0086, +0.0040]	81.5
Direct	Description	Soft prompting	0.5739	0.5733	-0.0006 [-0.0119, +0.0108]	40.0
Direct	Description	X-only	0.5739	0.5670	-0.0069 [-0.0201, +0.0056]	50.0
Curriculum	Reaction	Swapped Z	77.0000	77.0000	+0.0000 [+0.0000, +0.0000]	99.0
Curriculum	Reaction	Mean Z	77.0000	77.0000	+0.0000 [+0.0000, +0.0000]	98.0
Curriculum	Reaction	Soft prompting	77.0000	78.0000	+1.0000 [-2.0000, +4.0000]	90.0
Curriculum	Reaction	X-only	77.0000	77.5000	+0.5000 [+0.0000, +1.5000]	94.0
Direct	Reaction	Swapped Z	77.0000	78.0000	+1.0000 [+0.0000, +2.5000]	99.0
Direct	Reaction	Mean Z	77.0000	78.0000	+1.0000 [+0.0000, +2.5000]	98.0
Direct	Reaction	Soft prompting	77.0000	77.5000	+0.5000 [-2.5000, +3.5000]	91.5
Direct	Reaction	X-only	77.0000	78.5000	+1.5000 [-1.0000, +4.0125]	92.5

C.2 DOES A CONTINUOUS BLOCK REMAIN NECESSARY?

Table 14 retains both removal and zeroing for the latent and soft-prompt models. Scores remain close to their respective baselines even when the block is absent. Together with their unchanged answer onset, this limits the generality of LatentChem's dependence on a prefix for response organization.

Table 14: Latent removal and zeroing in all four controlled models. Description scores are METEOR; reaction scores are exact-match percentages. Every listed condition has 100% strict answer-tag onset and no detected explicit steps.

	Description			Reaction		
Model	Matched	X-only	Zero Z	Matched	X-only	Zero Z
Curr. / Latent reasoning	0.5727	0.5777	0.5694	77.0	77.5	78.0
Curr. / Soft prompting	0.5769	0.5739	0.5728	78.0	78.0	78.5
Direct / Latent reasoning	0.5739	0.5670	0.5751	77.0	78.5	78.0
Direct / Soft prompting	0.5733	0.5723	0.5745	77.5	78.0	78.0

D REPRESENTATION ANALYSES

The analyses distinguish recoverable molecular information from the magnitude and structure of input-dependent differences. Probe controls, latent geometry, and attention are reported separately because they do not establish the same claims.

D.1 SIZE-CONTROLLED PROBES

Latent features average the ten LatentChem positions or eight controlled-model positions. LatentChem's input reference averages its 128 initial molecular projection queries. Molecular weight (MW), logP, topological polar surface area (TPSA), and ring count are computed with RDKit on the largest molecular fragment. Size covariates are SMILES token count and heavy-atom count. Description probes use 200 molecules and five outer scaffold-grouped folds, with three inner folds selecting ridge regularization from $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. The same molecules and folds are used across models. Empty Murcko scaffolds are grouped together.

Feature standardization, division by the square root of feature dimension, size-effect fitting, and regularization selection use training folds only. Linear size effects are regressed out of both representations and targets. The predicted size component is added back to held-out residual predictions; ΔR^2 subtracts the size-only score on the same observations. Quadratic controls add both squared size covariates and their interaction. Intervals use 1,000 paired molecule-bootstrap samples of out-of-fold predictions, conditional on the fitted cross-validation models. Shuffled-target baselines rerun nested selection for 100 permutations.

The structural-label analysis starts from 85 RDKit `fr_` labels, binarized for presence. Thirty labels have at least five positive and five negative examples in every outer and inner training subset. This support rule does not depend on prediction scores. Each label uses a ridge classifier, with inner-fold average precision selecting regularization. Macro AP averages across labels; bootstrap samples retain each molecule's full label vector. Three duplicate label vectors leave 27 distinct labels; deduplicated AP is 0.3495 for size alone, 0.7126 for H_0, and 0.5690 for Z.

Table 15: Size-controlled property readout on ChEBI-20. Intervals resample held-out predictions at the molecule level.

Representation / property	Size-only R^2	Combined R^2	ΔR^2 [95% CI]
Z mean / MW	0.9838	0.9834	-0.0004 [-0.0017, +0.0004]
Z mean / logP	0.6393	0.7833	+0.1440 [+0.0473, +0.2575]
Z mean / TPSA	0.8189	0.9083	+0.0894 [+0.0519, +0.1580]
Z mean / Rings	0.4715	0.8628	+0.3913 [+0.3012, +0.5163]
H0 mean / MW	0.9838	0.9839	+0.0001 [-0.0020, +0.0019]
H0 mean / logP	0.6393	0.8514	+0.2121 [+0.1018, +0.3530]
H0 mean / TPSA	0.8189	0.9421	+0.1232 [+0.0792, +0.1984]
H0 mean / Rings	0.4715	0.9387	+0.4672 [+0.3714, +0.6005]

D.2 VARIANCE, DISPERSION, AND SWAP AMPLITUDE

For n molecules and T latent steps, let $\mu = (nT)^{-1} \sum_{i,t} z_{it}$ be the grand mean of $z_{it} \in \mathbb{R}^d$. The decomposition uses $a_i = T^{-1} \sum_t z_{it} - \mu$, $b_t = n^{-1} \sum_i z_{it} - \mu$, and $r_{it} = z_{it} - \mu - a_i - b_t$. Here a_i is the molecule effect, b_t the shared step effect, and r_{it} their interaction. Orthogonality gives

$$\sum_{i,t} \|z_{it} - \mu\|^2 = T \sum_i \|a_i\|^2 + n \sum_t \|b_t\|^2 + \sum_{i,t} \|r_{it}\|^2. \tag{3}$$

Each term is normalized by total centered squared variation.

Table 16: Linear and quadratic size controls on the same 200 description molecules and scaffold folds. Cells show ΔR^2 with 95% intervals where available; LC denotes LatentChem, and Curr. and Direct denote controlled latent-reasoning models.

Model	Control	MW	logP	TPSA	Rings
LC	linear	-0.0004	+0.1440 [+0.0473, +0.2575]	+0.0894 [+0.0519, +0.1580]	+0.3913 [+0.3012, +0.5163]
LC	quadratic	-0.0004	+0.1376 [+0.0600, +0.2400]	+0.1055 [+0.0488, +0.2191]	+0.3660 [+0.2102, +0.6414]
Curr.	linear	+0.0022	+0.2291 [+0.1201, +0.3680]	+0.1304 [+0.0846, +0.2072]	+0.4577 [+0.3661, +0.5855]
Curr.	quadratic	+0.0028	+0.1950 [+0.1245, +0.2870]	+0.1370 [+0.0817, +0.2486]	+0.4614 [+0.2811, +0.8030]
Direct	linear	+0.0012	+0.2315 [+0.1144, +0.3702]	+0.1266 [+0.0804, +0.2014]	+0.4564 [+0.3633, +0.5868]
Direct	quadratic	+0.0014	+0.2212 [+0.1234, +0.3480]	+0.1446 [+0.0794, +0.2816]	+0.4667 [+0.2736, +0.8312]

Table 17: Readout of 30 functional-group and local-structure labels on ChEBI-20. Macro average precision (AP) and shuffled-label baselines use the same label set.

Representation	Macro AP [95% CI]	Shuffled AP
Size only	0.3417 [0.3240, 0.3735]	0.2516
H0 mean	0.7119 [0.6821, 0.7520]	0.2743
Z mean	0.5710 [0.5476, 0.6111]	0.2729

Relative dispersion at step t is $\mathrm{RMS}_i(z_{it} - \bar{z}_t)/\,\mathrm{RMS}_i(z_{it})$, including coordinates in both RMS averages. For the actual source–donor pair A–B,

$$\alpha_{A,B} = \left[\frac{1}{T}\sum_t \frac{\mathrm{RMS}(z_{Bt} - z_{At})^2}{\mathrm{RMS}(z_{At})^2}\right]^{1/2}. \tag{4}$$

This measures perturbation magnitude in the same units as the noise scan; donor differences retain their native directions. All 200 LatentChem description swaps and 600 LatentChem optimization swaps have $\alpha \leq 0.25$. The median is 0.0767 for description and 0.0384–0.0464 across optimization tasks. LatentChem contracts early in the trajectory, while the controlled-model description dispersions increase across positions.

Table 18: Centered latent variance by source on the official test cohorts (200 description molecules, 100 inputs per optimization task, and 200 controlled-model reaction inputs). The molecule and interaction terms are input-dependent; their sum ranges from 1.151% to 1.417% across the six optimization tasks. The step term is shared across inputs. Static prompts have zero molecule and interaction terms by construction.

Model	Task	Molecule (%)	Shared step (%)	Interaction (%)
LatentChem	Description	0.799	96.234	2.966
LatentChem	QED	0.242	98.592	1.166
LatentChem	LogP	0.305	98.691	1.005
LatentChem	Solubility	0.198	98.849	0.953
LatentChem	DRD2	0.230	98.609	1.161
LatentChem	JNK3	0.202	98.839	0.959
LatentChem	GSK3-β	0.249	98.583	1.168
Curr. / Latent reasoning	Description	13.638	82.508	3.853
Curr. / Latent reasoning	Reaction	12.537	84.099	3.364
Direct / Latent reasoning	Description	23.212	70.719	6.069
Direct / Latent reasoning	Reaction	28.030	65.531	6.439

Table 19: RMS-normalized amplitudes of actual substitutions. Intervals are reported where available; LatentChem description identity is 16.5%.

Model	Task	Median α [95% CI]	Range	Identity
LatentChem	Description	0.0767	0.0336–0.2248	16.5%
LatentChem	QED	0.0435	0.0248–0.1582	93.0%
LatentChem	LogP	0.0448	0.0266–0.1178	92.0%
LatentChem	Solubility	0.0384	0.0218–0.2058	90.0%
LatentChem	DRD2	0.0464	0.0198–0.1467	90.0%
LatentChem	JNK3	0.0401	0.0198–0.1296	87.0%
LatentChem	GSK3-β	0.0431	0.0182–0.1205	93.0%
Curr. / Latent reasoning	Description	0.2080 [0.2032, 0.2112]	0.1266–0.3264	75.0%
Curr. / Latent reasoning	Reaction	0.1999 [0.1965, 0.2041]	0.1294–0.2910	99.0%
Direct / Latent reasoning	Description	0.2913 [0.2805, 0.2985]	0.1831–0.4117	79.5%
Direct / Latent reasoning	Reaction	0.3298 [0.3223, 0.3380]	0.2029–0.4468	99.0%

Table 20: Between-molecule dispersion at the first, fourth, and final latent positions. The final position is 10 for LatentChem and 8 for controlled models. Showing the final value also captures the rebound after LatentChem's early contraction.

Model	Task	Position 1	Position 4	Final
LatentChem	Description	0.1751	0.0229	0.0418
LatentChem	DRD2	0.1078	0.0142	0.0263
LatentChem	LogP	0.1038	0.0168	0.0304
LatentChem	QED	0.1078	0.0141	0.0283
LatentChem	Solubility	0.0957	0.0132	0.0293
LatentChem	GSK3-β	0.1073	0.0148	0.0240
LatentChem	JNK3	0.0964	0.0139	0.0225
Curr. / Latent reasoning	Description	0.1173	0.1484	0.1803
Curr. / Latent reasoning	Reaction	0.1312	0.1403	0.1558
Direct / Latent reasoning	Description	0.1670	0.2026	0.2473
Direct / Latent reasoning	Reaction	0.2238	0.2301	0.2510

D.3 ATTENTION ANALYSIS

The chemical-keyword rule lowercases each token string, removes tokenizer whitespace markers, and tests for a substring match against the following fixed list: acid, alcohol, aldehyde, alkane, alkene, alkyne, amide, amine, aromatic, benzene, bond, carbonyl, carboxyl, ester, ether, halogen, hydroxyl, ketone, methyl, nitrogen, oxygen, phenyl, phosphate, ring, salt, stereoisomer, sulfide, and sulfonamide. This is a token-string heuristic, not a semantic annotation of chemical claims.

Controlled attention analyses use the same 200 official-test description molecules and teacher-forced reference descriptions, capped at 96 reference tokens. For every token identified by the saved chemical-keyword rule, the query immediately preceding it is the position predicting that token. It is paired with the nearest non-chemical query in the same molecule and zero-based answer-token position bin: $[0, 4)$, $[4, 16)$, $[16, 32)$, $[32, 64)$, or $[64, 128)$. Ties choose the earlier position. For each query, attention weights are averaged across heads within each layer, summed over the latent positions, and averaged equally across all 32 layers. We average the chemical-minus-control attention mass assigned to Z within each molecule, then weight molecules equally. The 153 molecules with eligible pairs in all four models form the common comparison. Intervals use 5,000 paired molecule-bootstrap resamples. Figure 4c reports the results. The summary concerns attention to Z. These attention differences are descriptive and do not establish a causal contribution to prediction (Jain & Wallace, 2019).

E COMPLETE QUALITATIVE TEST EXAMPLE

The following saved test-set responses underlie Figure 1. Text is reproduced through the first completion boundary; chemical errors and original answer tags are retained. These are verbatim model outputs, not corrected chemical descriptions; this example illustrates response organization rather than estimating the frequency of any error.

Input SMILES: `CSCC[C@@H]1C(=O)NC(=O)N1`. The input is ChEBI-20 test sample 114629; its swapped latent comes from test sample 4124016 (the orange structure in Figure 1). Donor SMILES: `CC(=O)NC1=CC(=C(C=C1)N=NC2=C(C=CC3=CC(=CC(=C32)O)S(=O)(=O)[O-])N)S(=O)(=O)[O-]`.

Matched Z and Swapped Z (identical complete responses)

> .<answer> The molecule with the SMILES string CSCC[C@@H]1C(=O)NC(=O)N1 is a thioether derivative containing a bicyclic imide core. The structure consists of a five-membered ring (1,3-diazinane-2,4-dione) with two carbonyl groups at positions 2 and 4, and a thioether group (C-S-C) attached to a carbon atom in the ring. The chiral center at the ring-attached carbon (C[C@@H]) indicates stereochemistry, with the thioether group occupying the beta position relative to the ring closure. The molecule exhibits both thioether and imide functional groups, contributing to its polarity and potential reactivity. </answer>

X-only

> Step 1: input_structure: CSCC[C@@H]1C(=O)NC(=O)N1
>
> Step 2: parse_structure: The molecule contains a thioether group (CSCC) attached to a chiral carbon, which is part of a 5-membered ring with two carbonyl groups and two nitrogen atoms (C(=O)NC(=O)N).
>
> Step 3: identify_functional_groups: Thioether (C-S-C), chiral center (C[C@@H]), cyclic urea derivative (C(=O)NC(=O)N in a 5-membered ring).
>
> Step 4: analyze_structure: The thioether group is a non-chiral substituent. The chiral carbon connects the thioether to the 5-membered ring. The ring contains two carbonyl groups and two nitrogen atoms, forming a cyclic urea structure.
>
> Step 5: <answer> Thioether-linked 5-membered cyclic urea with a chiral center </answer>

X-only + hard prompt

> _id: 128001359
>
> Step 1: input_structure: CSCC[C@@H]1C(=O)NC(=O)N1
>
> Step 2: molecule_analysis: The molecule contains a thioether group (C-S-C) attached to a chiral carbon, which is part of a 5-membered ring. The ring consists of two carbonyl groups (C=O) and two amide nitrogens (N-C=O).
>
> Step 3: functional_group_identification: Thioether (C-S-C), chiral center (C[C@@H]), cyclic urea-like structure (C(=O)NC(=O)N).
>
> Step 4: structural_analysis: The thioether chain connects to a chiral carbon that forms a 5-membered ring with two carbonyl groups and two amide nitrogens, creating a cyclic urea derivative.
>
> <answer> A molecule with a thioether group (C-S-C) attached to a chiral carbon, which is part of a 5-membered ring containing two carbonyl groups and two amide nitrogens. </answer>