SMOLLM: SMALL LANGUAGE MODELS LEARN SMALL MOLECULAR GRAMMAR

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models for molecular design have scaled to hundreds of millions of parameters, yet how they learn chemical grammar is poorly understood. We train SMolLM, a 53K-parameter weight-shared transformer, to generate novel SMILES with 95% validity on the ZINC-250K drug-like-molecule benchmark, outperforming a standard GPT with 10 times more parameters. Mechanistically, the same block resolves SMILES constraints across passes in a fixed hierarchy: brackets first, rings second, and valence last, as shown by error classification and linear probing, with ablation isolating the bracket-matching head. Together, these results yield a compact, mechanistically interpretable molecular generator and a testbed for studying iterative computation in formal-language domains.

1 INTRODUCTION

Designing new molecules is a core task in drug discovery and materials science, and language models have become a major tool for molecular generation. From RNNs (Segler et al., 2018; Olivecrona et al., 2017) to transformer decoders (Bagal et al., 2022; Ross et al., 2025; Xu et al., 2026), these generators have grown by orders of magnitude in parameter count. However, whether such scale is necessary and what computations underlie valid generation remain poorly understood.

One promising way to reduce parameter count without reducing computational depth is to reuse the same transformer block repeatedly. Weight-shared transformers instantiate this idea by applying a single block iteratively for K passes. This matches the computational depth of a K-layer unshared stack with one block's worth of parameters. Such architectures are Turing-complete (Giannou et al., 2023), can run iterative algorithms like gradient descent (von Oswald et al., 2023), and match or exceed standard transformers on algorithmic reasoning (Dehghani et al., 2019), language modeling (Bae et al., 2025; Reid et al., 2021), and vision (Goyal et al., 2026). This makes them particularly suited to formal-language domains like SMILES, where validity depends on repeated local and long-range checks.

Weight sharing also makes generation interpretable by attributing behavior to computation rather than parameters. Prior work in mechanistic interpretability has decomposed circuits in standard transformers (Elhage et al., 2021; Olsson et al., 2022) and in molecular transformers (Varadi et al., 2025). In these architectures, however, each layer has its own weights, so functional differences between layers can reflect differences in parameters rather than differences in computation.

Here, we show that a single 53K-parameter weight-shared transformer block, trained on ZINC-250K (Irwin et al., 2012) SMILES strings (Simplified Molecular Input Line Entry System), generates novel SMILES with 95% validity. At 53K parameters, the model is over an order of magnitude smaller than the smallest prior transformer-based molecular generator we compare against (Bagal et al., 2022; Ross et al., 2025; Xu et al., 2026), and outperforms an identically-trained 527K-parameter unshared GPT in validity (95.3% vs. 87.6%) using 10 times fewer parameters.

Beyond generating valid molecules, the block is mechanistically interpretable. With weight sharing, the same parameters run at every pass, so functional differences across passes reflect changes in computation, not changes in weights. SMILES validity depends on three hard constraints: brackets must match, ring digits must pair at long range, and atom valences must be chemically valid. Because each constraint has an exact symbolic check, we can measure how generation improves at every intermediate pass.

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.

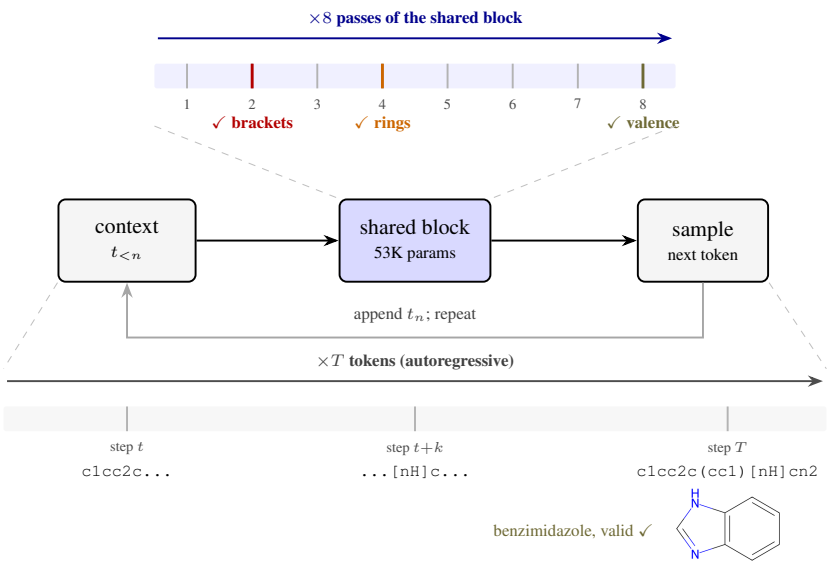


Figure 1: **Overview of SMolLM.** *Top track:* for each emitted token, the same 53K shared block runs eight passes. By truncating inference at intermediate passes, we find that the shared block solves grammar in stages, with each stage retained as depth increases: brackets by pass 2, rings by pass 4, valence by pass 8. *Bottom track:* the model autoregressively emits T SMILES tokens. The molecule at step T is benzimidazole, which is generated by WS-53K-s42 in a 10,000-sample run and absent from the training data.

Using three analyses (error classification across passes, a combined probing/sparse-autoencoder study of internal representations, and ablation), we show that the model learns chemical grammar pass by pass: brackets are resolved by pass 2, rings by pass 4, and valence by pass 8 (Fig. 1). Sparse autoencoders and linear probing recover the same ordering in the internal representations, from individual tokens in early passes to bracket depth and ring state in middle passes (ring-state accuracy climbs from 79% at pass 1 to 98% by pass 5). Ablation localizes bracket matching to a single attention head in WS-206K, where ablating it at pass 1 raises bracket errors by 23 percentage points without changing ring or valence errors, while the smaller WS-53K distributes the same function across heads.

Together, these results show a shared block resolving constraints from local to global scope, suggesting that similar pass-wise structure may appear in other formal-language domains with hierarchical constraints (Section 4). The weight-shared block thus serves as both a compact generator for molecular design and a small testbed for iterative computation.

2 Experimental Setup

This section defines the SMILES generation task, the weight-shared and baseline models, and the training and evaluation protocol.

SMILES. SMILES is a text encoding of molecular graphs. For example, atoms can be written with symbols such as `C`, `N`, and `O`. Bonds can be implicit or written explicitly with symbols such as `=` and `#`. Branches use parentheses, and ring closures use paired digits. The string `CC(=O)O`, for instance, encodes acetic acid. A generated string is valid only if brackets balance, ring digits pair, and atom valences are chemically possible. Validity rate is the fraction of generated strings that decode to real molecules via RDKit. We use SMILES rather than SELFIES (Krenn et al., 2020) because SMILES leaves these grammar and chemistry constraints for the model to learn, whereas SELFIES guarantees validity by construction.

Architecture. All models are autoregressive transformers based on GPT (Radford et al., 2019). We compare three variants. Standard GPT uses learned positional embeddings and Gaussian Error Linear

Unit (GELU) activations with a feed-forward expansion factor of 4. Modern GPT replaces these with rotary position embeddings (RoPE; Su et al., 2024) and Swish-Gated Linear Units (SwiGLU; Shazeer, 2020). Weight-Shared (WS) reuses a single Modern GPT block for D forward passes, equivalent to a looped transformer (Giannou et al., 2023). We call D its *virtual depth*. All transformer variants apply pre-norm LayerNorm (LayerNorm before each sublayer), tie input/output embeddings, use dropout 0.1, and have no bias terms. We additionally evaluate Gated Recurrent Unit (GRU) baselines at parameter counts matched to WS-53K and WS-206K. Model names combine class and parameter count, so WS-53K denotes a weight-shared block with 53K parameters and default virtual depth $D{=}8$. GPT-527K and GPT-527K-mod denote unshared Standard and Modern GPTs, and GRU-53K denotes the GRU baseline. When a WS model uses a virtual depth other than the default $D{=}8$, its name includes a suffix such as `-D4` or `-D16`. WS-53K and WS-206K are the two weight-shared blocks that recur, with WS-206K depth variants for the virtual-depth sweep. Full configurations are in Appendix A.

Training. We train on ZINC-250K (Irwin et al., 2012), the standard benchmark of 249,455 drug-like molecules for SMILES-based molecular generation (Bagal et al., 2022; Ross et al., 2025), with SMILES augmentation at $n_{\text{aug}}{=}10$ (Bjerrum, 2017) and a character-level tokenizer (vocab 50, max length 128). We optimize with AdamW for 100 epochs at batch size 256, using a cosine learning-rate schedule (peak 5×10^{-4}, 200-step warmup), mixed-precision training (AMP), and gradient clipping at 1.0. Core experiments use 3 seeds (42, 43, 44); supporting ablations use 1 seed. The reported experiments used approximately 500 GPU-hours on AMD Instinct MI210/MI250 GPUs. Distillation and DPO settings are given in Appendix G.

Evaluation. We sample $N{=}1{,}000$ molecules per model with temperature 0.8 and top-$k{=}40$. We report RDKit validity, Fréchet ChemNet Distance (FCD; Preuer et al., 2018), internal diversity (1 minus mean pairwise Tanimoto similarity, a Jaccard-style overlap measure, computed on Morgan radius-2 1024-bit fingerprints), uniqueness, and novelty, defined as the fraction of generated molecules absent from the ZINC-250K training set. We categorize invalid SMILES by failure mode for the error classification analysis (Section 4.1). Further details are in Appendix B.

3 Compact molecular generation

A 53K-parameter weight-shared transformer reaches 95.3% SMILES validity by sharing a single block across 8 passes, outperforming an unshared GPT with 10 times more parameters and approaching the validity of a GPT with 60 times more parameters.

3.1 Pareto frontier

We first ask whether weight sharing improves the tradeoff between parameter count and SMILES validity. We show that weight-shared models dominate the sub-megaparameter Pareto frontier. At every budget below 1M, the weight-shared curve delivers higher validity at fewer parameters (Fig. 2; full numbers in Appendix A, Table 5). WS-53K outperforms GPT-527K by 7.7 points (95.3% vs. 87.6%) and matches its internal diversity (0.852 vs. 0.851), at 10 times fewer parameters. The generated molecules show little evidence of memorization: across 3 seeds, 0.13% of valid WS-53K samples are exact training copies, and the mean nearest-neighbor Tanimoto similarity to training molecules is 0.52, below the nearest-neighbor similarity within the training set (0.59). WS-53K also matches GPT-3.2M to within 5 points (95.3% vs. 99.4%) at 60 times fewer parameters. FCD, a distributional-similarity metric where lower is better, is modestly higher for weight-shared models (2.5–2.9 vs. 2.0–2.4), trading some distributional fidelity for parameter efficiency. Thus, weight sharing shifts the compact-generation frontier rather than merely matching smaller baselines.

3.2 Virtual depth

We vary how many times the shared block is applied, measuring whether validity improves over repeated passes. Validity rises sharply with virtual depth and plateaus at $D{=}8$: 33% $\rightarrow$ 91% $\rightarrow$ 97% $\rightarrow$ 99% $\rightarrow$ 99% across $D \in \{1, 2, 4, 8, 16\}$ (WS-206K, 3 seeds per cell). The $D{=}1$ value ($33.0 \pm 2.3\%$, 3 seeds, 1,000 samples) matches the separately-trained WS-206K-D1 row of Table 1

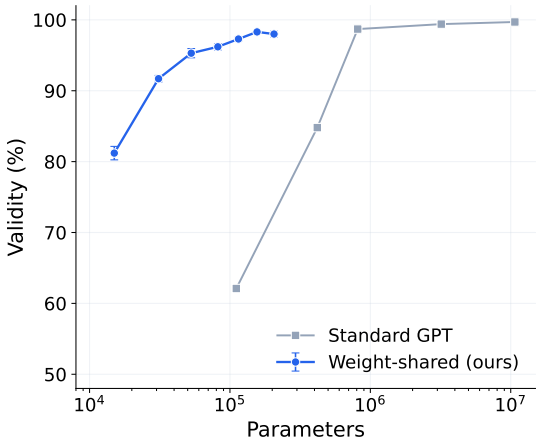


Figure 2: Pareto frontier. Weight-shared models dominate unshared GPTs below 1M parameters.

(33.0 $\pm$ 2.3%). The 58-point rise from D=1 to D=2, followed by diminishing returns through D=8 and a plateau at D=16, suggests that repeated passes iteratively refine validity.

We next isolate whether the gain comes from iteration itself rather than from other block-level architectural choices. An unshared Standard GPT baseline with four attention heads reaches 91.6 $\pm$ 0.3% validity. The matched Modern GPT adds RoPE and SwiGLU and reaches 94.4 $\pm$ 0.2%. The WS model reuses that Modern block for 8 passes and reaches 98.0% validity, adding 3.6 points while using half the parameters of the unshared baselines (Appendix A, Table 6), consistent with the Pareto dominance of looped architectures at larger scale (Reid et al., 2021; Prairie et al., 2026; Zeitoun et al., 2026).

To test whether iteration explains the gain, we compare against three alternatives. First, removing iteration from WS-206K collapses validity by 65 points. Second, larger unshared GPTs remain below WS-206K. GPT-241K reaches 88.4 $\pm$ 0.6%, and GPT-228K-mod reaches 91.2 $\pm$ 1.4% (3 seeds each), despite having more parameters. Third, extra training compute does not close the gap. A matched Modern GPT reaches 94.2 $\pm$ 0.1% after 800 epochs (8 times longer, with roughly twice WS-206K's training compute), similar to its 100-epoch value and still 3.8 points below WS-206K. Thus, the compact model succeeds by refining SMILES validity over repeated applications of the same small block.

3.3 Robustness checks

We test whether this compact-generation result survives changes in model class, SMILES augmentation, sampling settings, and post-training.

First, we compare against recurrent models at matched parameter count. GRU controls trail the weight-shared block, with GRU-53K reaching 87.7% and GRU-206K reaching 95.5% (2 seeds each, 25–30 training epochs vs. 100 for the WS comparators; matched-epoch comparison left to future work). At matched parameter count, WS exceeds the GRU by 7.7 points at 53K and 2.5 points at 206K, narrowing with capacity in line with larger GRUs (>1M) reaching high validity (Segler et al., 2018). Thus, the GRU comparison supports the compact-generation result at matched parameter count, while showing that the advantage narrows as recurrent capacity increases.

Second, we reduce SMILES enumeration by training with one string representation per molecule instead of 10 randomized SMILES. Under this setting, WS-53K drops from 95.3% to 92.3%, a 3.0-point decrease, and WS-206K drops from 98.0% to 97.8%, a 0.2-point decrease within seed variance (3 seeds), so augmentation helps but less than architecture.

Third, we vary decoding temperature, top-k, and top-p. Across temperature ($T \in \{0.7, 0.8, 1.0, 1.2\}$), top-$k$ ($\{40, \infty\}$), and top-p ($\{0.95, \text{none}\}$), WS-206K remains above the 403K unshared Modern GPT in every effective cell despite using half as many parameters (sign test $p<10^{-3}$ over 12 effective cells, since at $T \leq 0.8$ the top-k=40 and top-k=∞ cells are bitwise-identical).

Table 1: Error classification for separately-trained WS-206K-DN models at depths $N \in \{1, 2, 4, 8, 16\}$, with 1,000 samples per depth over 3 seeds ($\pm$ is standard deviation). Each invalid molecule is assigned to one failure class (Brackets, Rings, Valence, Other), rather than counting all failures in the string.

Passes	Valid (%)	Brackets (%)	Rings (%)	Valence (%)	Other (%)
1	33.0 ± 2.3	23.0 ± 1.7	12.8 ± 1.4	31.0 ± 1.3	0.3 ± 0.2
2	90.5 ± 1.8	1.3 ± 0.5	1.9 ± 0.3	6.2 ± 1.2	0.0 ± 0.1
4	97.1 ± 0.6	0.1 ± 0.1	0.8 ± 0.3	2.0 ± 0.9	0.1 ± 0.1
8	98.6 ± 0.2	0.1 ± 0.1	0.3 ± 0.2	1.0 ± 0.3	0.0 ± 0.1
16	98.7 ± 0.3	0.0 ± 0.1	0.5 ± 0.1	0.7 ± 0.3	0.1 ± 0.1

Finally, we test standard post-training methods. Offline knowledge distillation (KD; Hinton et al., 2015) drops WS-206K validity by $\sim$8 percentage points (3 seeds); the KL on soft labels spreads probability mass onto invalid tokens (Appendix G), and teacher quality has no effect (Appendix H). Direct Preference Optimization (DPO; Rafailov et al., 2023) preserves validity to within 1 pp (WS-206K $98.5 \rightarrow 97.6\%$; full results in Appendix I), so the compact-generation result is robust to preference fine-tuning. Together, these robustness checks support the interpretation that compact validity comes from repeated computation through the shared block, not from the sampler, memorization, or post-training.

4 Computation across passes

We next ask *how* weight-shared models generate valid molecules across repeated passes. Reusing a block provides iterative computation, but it does not determine whether successive passes perform distinct computations. We find here a fixed hierarchy of increasing scope. The shared block resolves brackets by pass 2 (local pairings within a few tokens), rings by pass 4 (long-range digit pairing across tens of tokens), and valence by pass 8 (atom-level integration of full bond context).

The rest of this section establishes this hierarchy with four complementary measurements: outputs, representations, causal structure, and capacity. We first classify errors across virtual depth to test whether failures disappear in the bracket-to-ring-to-valence order (Section 4.1). We then use linear probes and sparse autoencoders to ask whether hidden states follow the same progression from token-level features to bracket depth, ring state, and atom identity (Section 4.2). We next ablate individual heads to localize bracket matching to a single attention head in WS-206K (Section 4.3). Finally, we compare model sizes to test whether smaller blocks resolve the same constraints slightly later and with less localized circuitry while preserving the order itself (Section 4.4).

4.1 Errors resolve in order

To test whether repeated passes specialize at the output level, we classify errors in invalid molecules generated at each virtual depth. This sweep isolates virtual depth by holding block size fixed and varying D in separately trained WS-206K-DN models. Later analyses hold $D{=}8$ and compare WS-53K with WS-206K. Specialization should not only make invalid samples rarer, but it should also shift failure modes from simpler syntax errors toward harder chemical constraints. The three error classes instantiate this order: bracket errors test parenthesis matching, ring errors test long-range digit pairing, and valence errors test context-sensitive chemical consistency. We train a separate model at each depth ($D \in \{1, 2, 4, 8, 16\}$), which avoids reading early-pass representations through an output head calibrated for pass 8.

Table 1 shows the predicted shift in assigned failure modes. At 1 pass, many samples fail before reaching later checks. By pass 2, bracket errors have collapsed from 23.0% to 1.3%, consistent with early resolution of parenthesis matching. Among the remaining assigned failures, ring-closure errors then drop between passes 2 and 4 (1.9% to 0.8%), resolving long-range digit pairing. Valence violations are the last substantial residual class, dropping from 2.0% to 1.0% between passes 4 and 8. The main error sweep therefore supports the predicted bracket-to-ring-to-valence sequence.

The main sweep compares separately trained models, so the ordering could in principle reflect training a different model at each depth rather than computation inside one model. To control for

this depth-training confound, we read the trained 8-pass model's output head at each pass (Table 7). This intermediate readout gives lower absolute validity, as expected when a pass-8 head is applied to earlier representations, but the ordering is unchanged, with brackets the first error type to drop below 5% (pass 3), rings the second (pass 5), and valence the third (pass 6). Thus, the intermediate readout rules out training depth as the source of the ordering.

4.2 Representations organize across passes

The output-level analysis shows when each constraint is satisfied in the generated molecules, and we test whether the same hierarchy appears inside the model's hidden states as the shared block is applied repeatedly. We use two complementary tests on the hidden states of the trained 8-pass model (pass k refers to the state after the k-th application of the block). First, linear probes ask whether bracket depth and ring state are decodable at each pass. Second, sparse autoencoders (SAEs) ask whether interpretable chemical features emerge without specifying a readout in advance. Table 2 summarizes both analyses; they show the same bracket-then-ring ordering.

Table 2: Representation organization across passes. Probing rows report peak linear-probe accuracy. SAE rows report the peak feature–property correlation r across all learned features. Per-pass curves appear in Fig. 5.

		WS-53K		WS-206K	
Method	Property	peak	pass	peak	pass
Probing	Bracket depth	98.4%	3	99.2%	3
	Ring state	98.0%	5	98.6%	4
SAE	Bracket detector	0.97	2	0.80	1
	Ring-digit detector	0.99	3	0.93	3
	Bracket depth	0.64	2, 8	0.55	4
	Atom identity	0.84	2	0.62	1
	Ring state	0.64	5	0.75	4

Linear probing. We train linear probes on each pass of the 8-pass model to test when bracket depth and ring state become decodable. In both models, probes decode bracket depth earlier than ring state. WS-53K reaches 98.4% bracket-depth accuracy at pass 3, before ring-state accuracy reaches 98.0% at pass 5. WS-206K reaches 99.2% bracket-depth accuracy at pass 3, before ring-state accuracy reaches 98.6% at pass 4. Although WS-53K starts 3–6 points lower than WS-206K at pass 1, both models reach comparable probe accuracy by pass 5, so the difference is timing rather than final decodability. The ordering is the same in a single trained model as in the output-level analysis.

Sparse autoencoders. Sparse autoencoders test the same progression without training a linear probe for each property. For each pass, an SAE learns sparse features from hidden states. We then ask which features track five chemical properties: `is_bracket`, `is_ring_digit`, `bracket_depth`, `ring_state`, and `is_atom` (Appendix E). The learned features separate by pass most clearly in WS-53K. Token-level detectors (`is_bracket`, `is_ring_digit`) appear early, at passes 2–3 with correlations r=0.97 and 0.99. Atom identity also achieves high correlation early (pass 2 in WS-53K, r=0.84), consistent with atom-type information being available from early passes. Two features require integrating information across multiple tokens: bracket *depth* (the count of unmatched open brackets at each position) and ring *state* (which ring digits are currently open). Ring-state correlation reaches its peak at pass 5 (r=0.64). Bracket-depth correlation is bimodal: near-equal peaks at passes 2 and 8 (r=0.63 and 0.64), separated by a dip to $r\approx$0.49. WS-206K shows weaker separation (Table 2; Fig. 5). Because the SAEs are not trained to predict these properties, the features appear as emergent decompositions rather than only as supervised readouts.

Moreover, feature–property correlations are stable across corpus size (Appendix E). When we scale the SAE corpus from 2,000 to 20,000 ZINC-250K SMILES, WS-53K feature–property correlations strengthen or remain within 0.01. Taken together, error classification, supervised probes, and the SAE support the same brackets-to-rings-to-valence hierarchy.

Table 3: Head-ablation sweep for WS-53K and WS-206K. Changes are percentage points (pp) from the unablated baseline (n=2,000 per condition). The bracket-matching head is identified per seed (WS-206K heads 1/2/0, WS-53K heads 3/1/3 across seeds 42/43/44) and shown separately. Other heads and passes are aggregated.

	WS-53K		WS-206K	
Ablation	Δbrackets (pp)	Δvalidity (pp)	Δbrackets (pp)	Δvalidity (pp)
Bracket head, pass 1	+10	−10	+23	−23
Bracket head, pass 2	+5	−6	+17	−17
Bracket head, pass 3	+2	−2	+5	−5
Bracket head, pass 4+	<2	<2	<2	<2
All other heads, all passes	<2	<2	<2	<2

4.3 THE BRACKET-MATCHING HEAD

In both error classification and internal representations, bracket matching is the first constraint to resolve. We further test whether that first step can be traced to a specific component of the shared block. If so, ablating one attention head at one pass should selectively raise bracket errors, rather than simply lowering validity whenever any part of the block is removed. We ablate a head by zeroing its contribution at one pass, then measuring the full error profile. We repeat this intervention across all four heads and all eight passes, for a total of 32 head-pass combinations, and generate n=2,000 molecules per condition for both models. Table 3 summarizes the resulting changes in bracket errors and validity.

Across the full sweep, we localize the bracket-matching computation to a single attention head in WS-206K. Ablating that head at pass 1 raises bracket errors by 23 points while ring and valence errors remain at baseline; the effect persists through passes 2–3 (increases of 17 and 5 points) and then vanishes by pass 4, and all other heads at every pass remain near baseline. The dependence is thus specific to one head rather than to early-pass computation in general. WS-53K does not concentrate bracket matching the same way: at seed 42, no single head dominates (the largest pass-1 bracket-error increase is 3 points), and the 10-point increase in the 3-seed mean is carried by different heads in different seeds (heads 3, 1, and 3 for seeds 42, 43, and 44). The smaller model therefore distributes bracket matching across heads, whereas the larger model localizes it to one.

We use two controls to test whether the effect is specific to one head rather than to early-pass ablation or activation magnitude in general. To rule out general early-pass sensitivity, we ablate all of pass 1 in WS-206K rather than the bracket head alone; this leaves bracket errors at baseline and drops validity by only 0.5 points, roughly 50 times smaller than the 23-point drop from the single-head ablation. To rule out magnitude-driven effects, we compare activation norms: in all three seeds, the bracket head has the *smallest* pass-1 norm, and ablating the *largest*-norm head does not produce a larger validity drop. Together, the controls show that bracket matching depends on this specific head.

4.4 SMALLER MODELS NEED MORE PASSES

To test whether the pass-by-pass hierarchy changes with capacity, we compare WS-53K and WS-206K across error classification and probing. We find that WS-53K follows the same sequence as WS-206K but reaches its resolution points about one pass later (Table 8): per-pass validity peaks at pass 8 rather than pass 7, and the ring-state probe peaks at pass 5 rather than pass 4. Bracket depth peaks at pass 3 in both models, so the shift is small and appears only in the longer-range constraints. The two models reach comparable final accuracies (95.3% vs. 98.0% validity, and 98.4% vs. 99.2% best bracket-probe accuracy). Capacity also changes how bracket matching is implemented: WS-206K localizes it to a single head (Section 4.3), whereas WS-53K distributes it across heads. Thus, lower capacity slightly delays when each constraint resolves and spreads bracket matching across more heads, but does not change the order in which the constraints resolve.

5 DISCUSSION

This paper shows that weight sharing can make molecular generation both compact and interpretable. With only 53K parameters, a shared-block transformer generates highly valid SMILES while preserving a visible sequence of validity computation. This improves the parameter-validity tradeoff, showing that small models can recover much of the validity of substantially larger unshared transformers.

We observe the same sequence across generated errors, hidden-state probes, and ablations: bracket matching first, ring closure second, and valence last. This consistency suggests that the shared block does more than reduce parameter count. By reusing one set of weights across passes, the model exposes successive stages of validity computation.

The ordering also links the empirical result to the grammar of SMILES. Bracket matching, ring closure, and valence are not arbitrary metrics; they form a progression from local syntax to long-range structure to chemically constrained validity. This connects the observed sequence to formal-language results in which fixed-depth transformers struggle with context-free and higher languages, while iterative depth supplies the recurrence such constraints require (Hahn, 2020; Delétang et al., 2023). Weight sharing makes this connection visible because the same parameters are applied repeatedly, so cross-pass changes track computation through time rather than differences between learned layers.

Limitations. We qualify the claims in three ways. First, the empirical scope is deliberately focused. We study ZINC-250K SMILES, so the ordering should be read as a claim about validity computation in this representation rather than about molecular structure independent of representation. Second, the mechanistic measurements are finite-sample estimates. The SAE analysis uses 20,000 SMILES ($\sim$770K tokens), so rare features may only emerge at larger scale, and the ablation sweep uses 2,000 molecules per condition, enough to identify the bracket-head effect but not to resolve very small effects from other heads. Third, weight sharing reduces parameters but not autoregressive inference cost, because FLOPs scale with pass count.

Broader impacts. Compact molecular generators can lower the compute cost of exploratory molecule design and mechanism analysis. The same capability could be misused to propose molecules with undesirable biological activity, so generated molecules should not be used for synthesis prioritization without toxicity, safety, and feasibility screening.

Future directions. These results point to broader tests of whether shared depth can expose structured computation in other scientific domains. One direction is to extend to other biological, chemical, and materials systems where similar hierarchies of constraints arise, such as RNA folding from local base pairing to long-range tertiary contacts. A second direction is to span spatial and temporal scales by letting each pass resolve a different scale, from atomic interactions to macroscale dynamics and mechanics, providing a single architecture for problems that today require stitching separate models together. A third is to design distillation objectives that preserve the iterative computation, for example by matching per-pass representations rather than only output distributions.

6 RELATED WORK

Molecular generation with language models. SMILES generation has progressed from RNN-based models (Segler et al., 2018; Olivecrona et al., 2017) to transformer decoders such as MolGPT (Bagal et al., 2022), GP-MoLFormer (Ross et al., 2025), and REINVENT 4 (Loeffler et al., 2024). Recent scaling work studies five molecular representations from 1M to 650M parameters (Xu et al., 2026). Its smallest model is 19 times larger than WS-53K. To our knowledge, no prior transformer work explores molecular generation below 1M parameters, the regime where reusing computation offers a direct alternative to scaling parameter count.

Weight sharing and looped transformers. Prior work uses weight sharing to compress transformers and to turn depth into repeated computation. Compression models such as ALBERT (Lan et al., 2020), Subformer (Reid et al., 2021), MobileLLM (Liu et al., 2024), and Relaxed Recursive Transformers (Bae et al., 2025) apply shared blocks at scales 1,000 to 10,000 times larger, where shared models match unshared baselines with fewer parameters (Prairie et al., 2026; Zeitoun et al.,

2026). DeLiCaTe (Yu & Börjesson, 2022) brings the same compression motive to SMILES property prediction. Other work uses repeated depth as computation, including the Universal Transformer (Dehghani et al., 2019), theoretical results on looped transformers (Giannou et al., 2023), and recent looped models with iteration-wise specialization (Yu et al., 2026; Goyal et al., 2026). We use the same architectural device for a different purpose. Because SMILES validity depends on repeated grammar checks, weight sharing makes those checks observable as computation through time.

Large looped language models make repeated computation central, but they leave individual iterations difficult to interpret. Models such as Huginn (Geiping et al., 2025; Lu et al., 2025) and Ouro (Zhu et al., 2025) operate at billion-parameter scale, where different iterations can perform similar computations (Blayney et al., 2026). Natural language also lacks the exact symbolic validity checks available in SMILES. Our setting combines a sub-100K bottleneck with constraints that can be measured at every pass, so the iteration is easier to separate into steps. This complements work on thermodynamic phases in looped transformers (Lam, 2026), formal-language constraints and circuit complexity (Hu et al., 2025), and concurrent studies of per-step specialization in latent reasoning models (Li et al., 2026; Kohli et al., 2026; Kaissis et al., 2026). Where those studies rely mainly on correlational traces, training stages, or trace decoding (Dilgren & Wiegreffe, 2026), SMILES lets us test each intermediate computation with symbolic checks and causal ablation.

Mechanistic interpretability. Mechanistic interpretability provides tools for finding circuits and features, but standard layerwise analyses make temporal ordering hard to separate from layer identity. We combine sparse autoencoders (Bricken et al., 2023; Cunningham et al., 2024), causal ablation (Vig et al., 2020), and probing (Conneau et al., 2018). Varadi et al. (2025) identify branch-balancing and ring-closure circuits in a 25M-parameter molecular transformer, establishing that chemically meaningful SMILES circuits can be recovered with SAE-based analysis. Weight sharing lets us test how related constraints resolve across repeated applications of the same block, rather than only where they appear across layers with separate weights. This distinction matters because Tenney et al. (2019) find a correlational syntax-to-semantics arc across BERT layers, while subsequent work questions such arcs under controls for layer position (Niu et al., 2022). Shared weights remove that layer-identity confound. We also use causal ablation because attention-based analysis of bracket matching can mislead (Wen et al., 2023), and we validate SAE features against symbolic chemical ground truth rather than auto-interpretability scores (Heap et al., 2025). This per-pass analysis complements SMI-TED's single-extraction-point SAE (Cohen et al., 2025) and TIDE's temporal SAEs for diffusion transformers (Huang et al., 2026).

7 Conclusion

SMolLM, a 53K-parameter weight-shared transformer, reaches 95.3% validity on ZINC-250K, outperforms an unshared GPT with 10 times more parameters, and shifts the sub-megaparameter Pareto frontier. The architecture is also mechanistically interpretable. Across generated errors, hidden-state probes, and ablations, the shared block resolves SMILES constraints in a fixed pass-by-pass hierarchy (brackets by pass 2, rings by pass 4, valence by pass 8), consistent from 53K to 206K parameters, with capacity changing how quickly each constraint resolves rather than the order itself. Taken together, weight sharing is not only a compression device for molecular generation but also a way to study iterative computation in domains with exact symbolic constraints.

AI use statement

Large language models were used as coding and writing assistants throughout this work, including for analysis scripts, experiment orchestration, drafting and editing text, and checking quantitative claims against experiment artifacts and cited references against their primary sources. The authors take full responsibility for the content of this paper.

References

Sangmin Bae, Adam Fisch, Hrayr Harutyunyan, Ziwei Ji, Seungyeon Kim, and Tal Schuster. Relaxed recursive transformers: Effective parameter sharing with layer-wise LoRA. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.20672.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9): 2064–2076, 2022. doi: 10.1021/acs.jcim.1c00600.

Esben Jannik Bjerrum. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*, 2017. URL https://arxiv.org/abs/1703.07076.

Hugh Blayney, Álvaro Arroyo, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, Michael M. Bronstein, and Xiaowen Dong. A mechanistic analysis of looped reasoning language models. *arXiv preprint arXiv:2604.11791*, 2026. URL https://arxiv.org/abs/2604.11791.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

Jaron Cohen, Alexander G. Hasson, and Sara Tanovic. Unveiling latent knowledge in chemistry language models through sparse autoencoders. *arXiv preprint arXiv:2512.08077*, 2025. URL https://arxiv.org/abs/2512.08077. NeurIPS 2025 Workshop on AI for Accelerated Materials Design.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, 2018.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024. URL https://arxiv.org/abs/2309.08600.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. Universal transformers. In *ICLR*, 2019. URL https://arxiv.org/abs/1807.03819.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *ICLR*, 2023.

Connor Dilgren and Sarah Wiegreffe. Are latent reasoning models easily interpretable? In *COLM*, 2026. URL https://arxiv.org/abs/2604.04902.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. In *NeurIPS*, 2025. URL https://arxiv.org/abs/2502.05171.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *ICML*, 2023. URL https://arxiv.org/abs/2301.13196.

Sahil Goyal, Swayam Agrawal, Gautham Govind Anil, Prateek Jain, Sujoy Paul, and Aditya Kusupati. ELT: Elastic looped transformers for visual generation. In *ECCV*, 2026. URL https://arxiv.org/abs/2604.09168.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Automated interpretability metrics do not distinguish trained and random transformers. *arXiv preprint arXiv:2501.17727*, 2025. URL https://arxiv.org/abs/2501.17727.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. NIPS 2014 Deep Learning Workshop.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *ACL*, 2025. URL https://arxiv.org/abs/2502.19249.

Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Fu-Yun Wang, Yuchi Wang, Renrui Zhang, Peng Gao, and Hongsheng Li. TIDE: Temporal-aware sparse autoencoders for interpretable diffusion transformers in image generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(1):435–443, 2026. doi: 10.1609/aaai.v40i1.37006. URL https://arxiv.org/abs/2503.07050.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52 (7):1757–1768, 2012. doi: 10.1021/ci3001277.

Georgios Kaissis, David Mildenberger, Juan Felipe Gomez, Martin J. Menten, and Eleni Triantafillou. Step-resolved data attribution for looped transformers. *arXiv preprint arXiv:2602.10097*, 2026. URL https://arxiv.org/abs/2602.10097.

Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers. In *COLM*, 2026. URL https://arxiv.org/abs/2604.07822.

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

Wai-Lun Lam. Energy-entropy regularization: The true power of minimal looped transformers. *arXiv preprint arXiv:2601.09588*, 2026. URL https://arxiv.org/abs/2601.09588.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020. URL https://arxiv.org/abs/1909.11942.

Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi Li, Jian Yang, Chenghua Lin, and Min Zhang. Dynamics within latent chain-of-thought: An empirical study of causal structure. In *ICML*, 2026. URL https://arxiv.org/abs/2602.08783.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *ICML*, 2024. URL https://arxiv.org/abs/2402.14905.

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. doi: 10.1186/s13321-024-00812-5.

Wenquan Lu, Yuechuan Yang, Kyle Lee, Yanshu Li, and Enqi Liu. Latent chain-of-thought? Decoding the depth-recurrent transformer. *arXiv preprint arXiv:2507.02199*, 2025. URL https://arxiv.org/abs/2507.02199.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. Does BERT rediscover a classical NLP pipeline? In *COLING*, pp. 3143–3153, 2022.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017. doi: 10.1186/s13321-017-0235-x.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models. *arXiv preprint arXiv:2604.12946*, 2026. URL https://arxiv.org/abs/2604.12946.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. Subformer: Exploring weight sharing for parameter efficiency in generative transformers. In *Findings of EMNLP*, 2021. URL https://arxiv.org/abs/2101.00234.

Jerret Ross, Brian Belgodere, Samuel C. Hoffman, Vijil Chenthamarakshan, Jiri Navratil, Youssef Mroueh, and Payel Das. GP-MoLFormer: A foundation model for molecular generation. *Digital Discovery*, 4(10):2684–2696, 2025. doi: 10.1039/d5dd00122f. URL https://arxiv.org/abs/2405.04912.

Marwin H. S. Segler, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018. doi: 10.1021/acscentsci.7b00512.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jianlin Su, Ahmed Murtadha, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL*, 2019.

Kristof Varadi, Mark Marosi, and Peter Antal. Circuits, features, and heuristics in molecular transformers. *arXiv preprint arXiv:2512.09757*, 2025. URL https://arxiv.org/abs/2512.09757.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*, 2020.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, 2023. URL https://arxiv.org/abs/2212.07677.

Kaiyue Wen, Yuchen Li, Bingbin Liu, and Andrej Risteski. Transformers are uninterpretable with myopic methods: A case study with bounded Dyck grammars. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2312.01429.

Dong Xu, Qihua Pan, Sisi Yuan, Jianqiang Li, Zexuan Zhu, and Junkai Ji. Unveiling scaling behaviors in molecular language models: Effects of model size, data, and representation. *arXiv preprint arXiv:2601.22757*, 2026. URL https://arxiv.org/abs/2601.22757.

Chengting Yu, Xiaobo Shu, Yadao Wang, Yizhen Zhang, Haoyi Wu, You Wu, Rujiao Long, Ziheng Chen, Yuchi Xu, Wenbo Su, and Bo Zheng. SpiralFormer: Looped transformers can learn hierarchical dependencies via multi-resolution recursion. *arXiv preprint arXiv:2602.11698*, 2026. URL https://arxiv.org/abs/2602.11698.

Yi Yu and Karl Börjesson. Chemical transformer compression for accelerating both training and inference of molecular modeling. *Machine Learning: Science and Technology*, 3(4):045009, 2022. doi: 10.1088/2632-2153/ac99ba. URL https://arxiv.org/abs/2205.07582.

Abbas Zeitoun, Lucas Torroba-Hennigen, and Yoon Kim. Hyperloop transformers. *arXiv preprint arXiv:2604.21254*, 2026. URL https://arxiv.org/abs/2604.21254.

Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling latent reasoning via looped language models. *arXiv preprint arXiv:2510.25741*, 2025. URL https://arxiv.org/abs/2510.25741.

APPENDIX

A ARCHITECTURE CONFIGURATIONS

Table 4: Architecture configurations, parameter counts, and reproducibility identifiers. *Name* is the canonical identifier used throughout the paper; *checkpoint ID* is the matching filename under `checkpoints/0504_paper/` in the checkpoint release and the corresponding training-run name (tag `0504_paper`); links are withheld for anonymous review. Class is the block family (Standard GPT, Modern GPT with RoPE+SwiGLU, weight-shared Modern GPT, or GRU). The `-mod` suffix on a GPT name marks the Modern block; the `-D`N suffix on a WS name marks a non-default virtual depth (default D=8); the `-8ep` suffix marks the 800-epoch training-compute control (default 100 epochs).

Name	Class	Checkpoint ID	L	H	E	D	Params
GPT-111K	Standard GPT	`L2H2E64`	2	2	64	–	111K
GPT-419K	Standard GPT	`L2H2E128`	2	2	128	–	419K
GPT-527K	Standard GPT	`L2H2E144`	2	2	144	–	527K
GPT-813K	Standard GPT	`L4H4E128`	4	4	128	–	813K
GPT-3.2M	Standard GPT	`gpt-3.2m`	4	4	256	–	3.2M
GPT-10.7M	Standard GPT	`L6H6E384`	6	6	384	–	10.7M
GPT-419K-h4	Standard GPT	`L2-E128-vanilla`	2	4	128	–	419K
GPT-241K	Standard GPT	`L2-E96-vanilla`	2	4	96	–	241K
GPT-403K-mod	Modern GPT	`L2-E128-modern`	2	4	128	–	403K
GPT-403K-mod-8ep	Modern GPT	`L2-E128-modern-8ep`	2	4	128	–	403K
GPT-228K-mod	Modern GPT	`L2-E96-modern`	2	4	96	–	228K
GPT-467K-mod	Modern GPT	`L6-E80-modern`	6	4	80	–	467K
WS-15K	Weight-shared	`ws-e32-x8`	1	4	32	8	15K
WS-31K	Weight-shared	`ws-e48-x8`	1	4	48	8	31K
WS-53K	Weight-shared	`ws-e64-x8`	1	4	64	8	53K
WS-82K	Weight-shared	`ws-e80-x8`	1	4	80	8	82K
WS-115K	Weight-shared	`ws-e96-x8`	1	4	96	8	115K
WS-156K	Weight-shared	`ws-e112-x8`	1	4	112	8	156K
WS-206K	Weight-shared	`ws-e128-x8`	1	4	128	8	206K
WS-206K-DN	Weight-shared	`ws-e128-x`N	1	4	128	$N \in \{1, 2, 4, 16\}$	206K
GRU-53K	GRU	`gru-53k`	–	–	64	–	54K
GRU-206K	GRU	`gru-206k`	–	–	105	–	207K

Table 5: Validity and FCD across the size sweep (companion to Fig. 2). Validity cells with $\pm$ are means over 3 seeds; cells without are single-seed.

Model	Params	Validity (%)	FCD
Standard GPT			
GPT-111K	111K	62.1	4.58
GPT-419K	419K	84.8 ± 0.7	3.03
GPT-527K	527K	87.6 ± 0.7	3.01
GPT-813K	813K	98.7 ± 0.2	2.44
GPT-3.2M	3.2M	99.4 ± 0.1	2.32
GPT-10.7M	10.7M	99.7	2.08
Weight-shared (ours)			
WS-15K	15K	81.2 ± 1.0	4.57
WS-31K	31K	91.7 ± 0.5	3.10
WS-53K	53K	95.3 ± 0.7	2.76
WS-82K	82K	96.2 ± 0.4	2.66
WS-115K	115K	97.3 ± 0.3	2.57
WS-156K	156K	98.3 ± 0.2	2.51
WS-206K	206K	98.0 ± 0.4	2.45

Table 6: Architecture ablation (block type varies; Standard = learned positional embeddings + GELU; Modern = RoPE + SwiGLU). Each row adds one architectural modification; the final row removes iteration to isolate its contribution. WS halves parameters yet delivers the largest jump. Baselines use H=4 heads to match the weight-shared block, isolating iteration from head count; this differs from Table 5's GPT-419K (H=2, same embedding dimension and parameter count).

Model	Params	Validity (%)	Δ (pp)
GPT-419K-h4 (Standard)	419K	91.6 ± 0.3	–
GPT-403K-mod (Modern)	403K	94.4 ± 0.3	+2.8
+ weight sharing (WS-206K)	206K	98.0 ± 0.4	+3.6
Capacity-matched unshared baselines ($\geq$ WS-206K's 206K)			
GPT-241K (Standard)	241K	88.4 ± 0.6	–
GPT-228K-mod (Modern)	228K	91.2 ± 1.4	–
Iteration removed (WS-206K-D1)	206K	33.0 ± 2.3	−65.0 (vs. WS-206K)

B Training and evaluation details

Dataset. ZINC-250K (Irwin et al., 2012): 249,455 drug-like molecules. SMILES augmentation (Bjerrum, 2017): each canonical SMILES expanded to $n_{\text{aug}} = 10$ randomized SMILES via RDKit. Character-level tokenizer, vocabulary size 50, max sequence length 128.

Training. 100 epochs, AdamW (weight decay 0.1), batch size 256, cosine LR (peak 5×10^{-4}, 200-step warmup), mixed-precision training (AMP), gradient clipping 1.0, no label smoothing. Distillation (Section 3.3): offline KD with softmax temperature $\tau = 2.0$ and hard/soft loss mixing coefficient $\alpha = 0.5$; teacher GPT-467K-mod (98.5%). DPO (Section 3.3): 5 epochs, lr $= 10^{-5}$, KL regularization strength $\beta = 0.1$, sample budget 10K molecules per base, pairing valid samples against invalid ones (so usable pairs are bounded by the invalid count). Full DPO results are in Appendix I.

Evaluation. $N = 1{,}000$ molecules per model, temperature $T = 0.8$, top-$k = 40$. Metrics: validity (RDKit parsing), FCD (Fréchet ChemNet Distance, analogous to FID; Preuer et al., 2018), internal diversity (1− mean pairwise Tanimoto, Morgan fingerprints radius 2, 1024 bits), uniqueness, novelty.

Compute. Reported runs used single AMD Instinct MI210 or MI250 GPUs (64–128GB HBM class) on an internal cluster. Training time ranged from under 2 GPU-hours for the smallest 53K models to under 24 GPU-hours for the largest reported models; most sub-megaparameter runs completed in a few GPU-hours. The reported experiments used approximately 500 GPU-hours total, excluding preliminary and failed runs that were not used in the paper. Checkpoints and logs require less than 1GB per run.

Pass indexing. The paper uses 1-indexed pass numbers (pass 1 is the first block application; pass 8 is the last). The released code uses 0-indexed positions (`pass=0..7`); to map paper to code, subtract 1.

Reproducibility. Code (MIT license) will be released upon publication; links are withheld for anonymous review. Each experiment maps to a numbered configuration in `experiments/smiles_gen/configs.py` and can be reproduced via `python -m experiments.smiles_gen.run -exp <N>`. Trained checkpoints (3 seeds per config, all 21 architectures) will be released under `checkpoints/0504_paper/`, and training runs are logged on Weights & Biases (tag `0504_paper`). Two stability notes: WS-206K's 3-seed std is 0.33 percentage points, and run-to-run variance across independent training runs is consistent with this. The four Table 6 checkpoints were regenerated and verified within 2 percentage points of reported values at N=10,000 samples.

C Error classification across passes

Companion plot to Table 1: error-type composition across virtual depths in separately-trained WS-206K-DN models. With n=1,000 samples per depth averaged over 3 seeds, the cross-pass trends are robust.

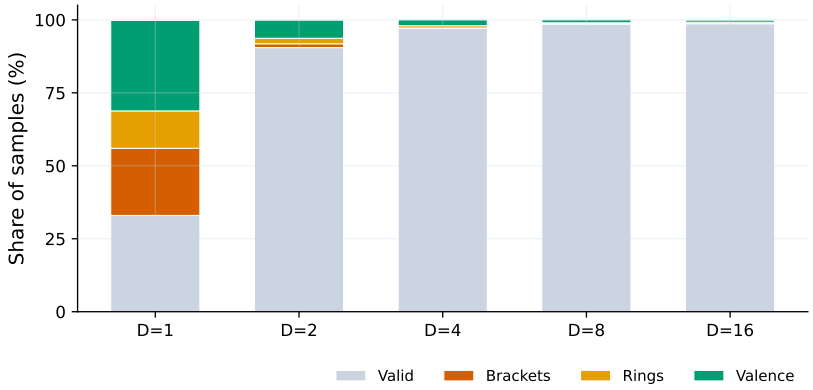


Figure 3: Bracket errors collapse by pass 2, rings by pass 4, valence by pass 8 (companion to Table 1).

Table 7: Per-pass readout from the trained 8-pass WS-206K model, averaged over 3 seeds (n=500 per pass per seed). Same constraint ordering as Table 1 (brackets→rings→valence); absolute validities are head-calibration-shifted. Valid reports the fraction of generated molecules that pass all checks. Brackets, Rings, Valence, and Other assign each invalid molecule to one failure class, rather than counting all possible failures in the same string.

Pass	Valid (%)	Brackets (%)	Rings (%)	Valence (%)	Other (%)
1	6.7	36.1	45.1	11.6	0.5
2	20.1	12.1	45.1	22.3	0.4
3	51.0	3.1	22.1	23.3	0.4
4	76.4	0.8	8.8	13.8	0.2
5	91.5	0.2	2.2	6.0	0.1
6	96.8	0.0	0.7	2.5	0.0
7	98.7	0.0	0.3	0.9	0.1
8	98.4	0.0	0.5	1.1	0.1

D HEAD-ABLATION HEATMAP

Companion to Table 3: full head × pass heatmap for both models. We apply a Bonferroni correction across the K=32 head-pass ablations for each model.

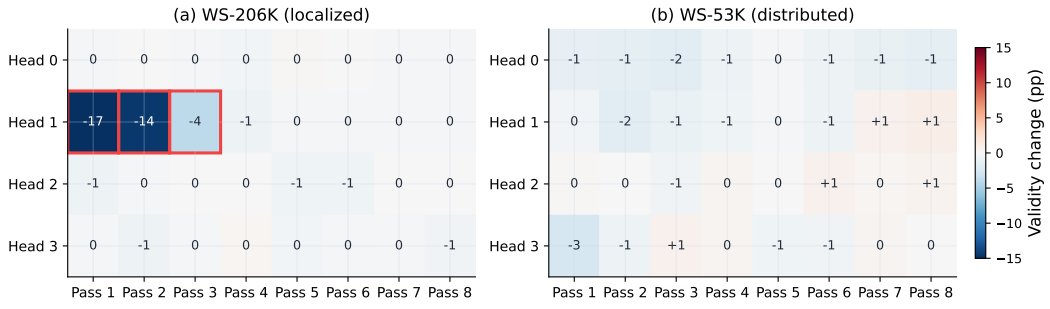
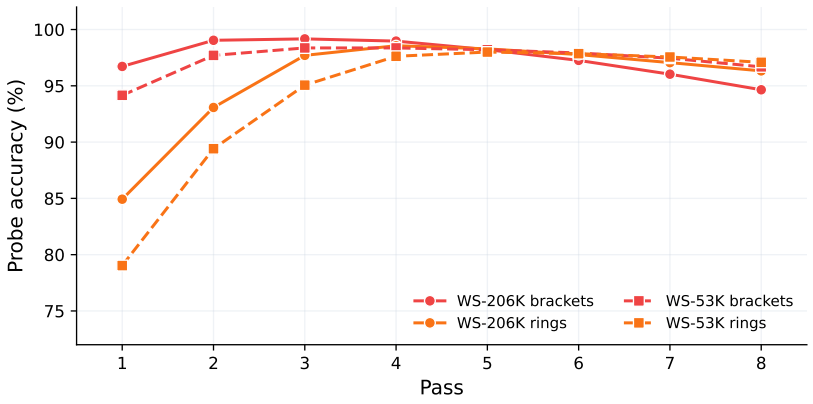


Figure 4: Ablation heatmap (head × pass, change in validity in percentage points; seed 42, n=2,000 per condition). WS-206K concentrates the damage in one head (head 1) at passes 1–3 and is negligible at every other head and pass. WS-53K shows no comparable single-head effect at this seed, consistent with bracket matching being distributed across heads in the smaller model (3-seed-mean bracket-head effect −10 pp; Table 3).

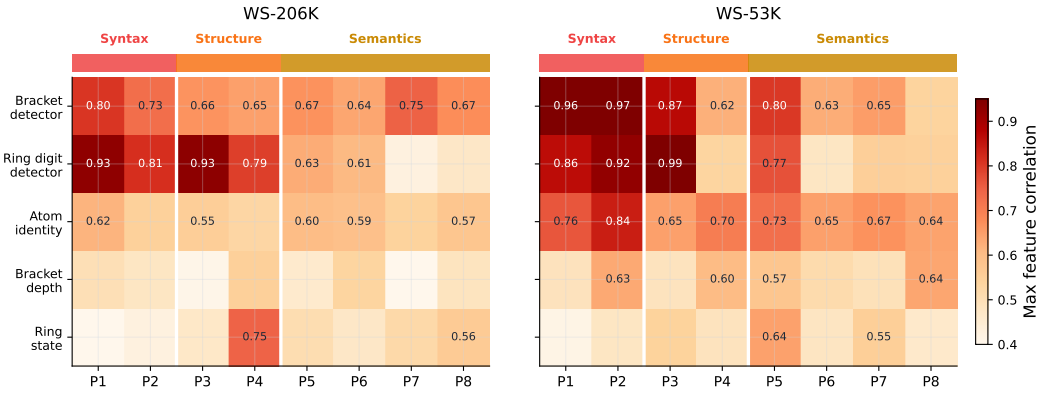
E REPRESENTATION ORGANIZATION ACROSS PASSES

Panel (a) shows linear-probe accuracy for bracket depth and ring state across passes. Panel (b) shows peak SAE feature correlation for five chemical properties. Companion to Table 2.

For each pass, we train an SAE with 512 features (4 times the hidden dimension for WS-206K and 8 times for WS-53K) using a sparsity penalty on 20,000 ZINC-250K SMILES (∼770K tokens). Decoder dictionary atoms are constrained to unit ℓ_2 norm following Bricken et al. (2023); the constraint reduces over the input dimension of the decoder weight so that each atom (column) has unit norm, which is necessary for cross-pass comparison of feature–property correlations under L1 sparsity. Scaling from 2,000 to 20,000 ZINC-250K SMILES increases or preserves the WS-53K feature–property correlations across all properties. The exact peak pass for an individual feature varies by ± 2 passes across seeds. For the late-pass convergence analysis, we compute cosine similarity between consecutive hidden states from the same trained 8-pass WS-206K model; pass 7–8 similarity reaches 0.995 while generated validity remains 97–98%.



(a) Bracket depth saturates 1–2 passes before ring state in both models.



(b) SAE cross-pass correlation heatmap (properties × passes) for WS-53K and WS-206K. Blank cells indicate peak feature correlation $r < 0.4$. All five properties reach $r \geq 0.6$ in at least one pass for WS-53K, and four of five for WS-206K.

Figure 5: Representation organization across passes. Panel (a) shows probe accuracy; panel (b) shows SAE feature correlation. Companion to Table 2.

F CAPACITY COMPARISON ACROSS METHODS

The diagnostics agree that WS-53K solves the same subtasks as WS-206K, with peaks appearing at the same pass or slightly later (Table 8). Five of the six diagnostic rows shift by 0 to +1 pass, with no negative shifts; the exception is the WS-53K SAE bracket-depth profile, which is bimodal

(near-equal peaks at passes 2 and 8) rather than single-peaked, spanning a -2 to $+4$ pass range relative to WS-206K. The overall picture is rate, not target: WS-53K reaches the same milestones on a slightly delayed schedule.

Table 8: Capacity comparison across diagnostics. "Shift" is WS-53K minus WS-206K peak pass (positive = WS-53K peaks later). Probing and validity peaks are 3-seed means; SAE peaks are single-seed estimates; ± 1-pass shifts are within probe noise.

Method	Measure	WS-53K	WS-206K	Shift
Validity	Per-pass validity peak	pass 8	pass 7	$+1$
Probing	Bracket depth peak	pass 3	pass 3	0
Probing	Ring state peak	pass 5	pass 4	$+1$
SAE	Bracket detector peak	pass 2	pass 1	$+1$
SAE	Bracket depth peak	passes 2, 8	pass 4	-2 / $+4$
SAE	Ring state peak	pass 5	pass 4	$+1$

G DISTILLATION ABLATION

An earlier version of this sweep used a label-smoothed teacher; smoothing made the soft-target signal nearly uniform and looked like evidence that KD helps. We re-ran the full sweep with a non-smoothed teacher (GPT-467K-mod; $\tau = 2.0$), and the apparent benefit disappeared: every mixture of soft targets underperforms hard-only (Table 9). The curriculum and curriculum+KD rows disambiguate the two effects: the scheduled data order alone gets 96.7% validity, and adding soft targets on top drops it to 90.3%, so the residual damage comes from soft targets themselves rather than from the training schedule.

Table 9: Distillation ablation (WS-206K, 1 seed except curriculum+KD which is 3 seeds). Hard-only and curriculum rows use the distillation pipeline and should be compared within this table, not against Table 5.

Condition	Alpha	Validity (%)	FCD
Hard-only	1.0	98.10	2.23
Curriculum	n/a	96.70	2.34
Mixed	0.5	91.90	2.31
Curriculum + KD	0.5	90.33 ± 0.49	2.45
Soft-only	0.0	89.30	2.21

H TEACHER QUALITY

Teacher validity spans $\sim$1 percentage point across the three candidates (GPT-3.2M 99.4%, GPT-467K-mod 98.5%, WS-206K self-distill 98.5%). Distillation from the two external teachers yields student validities within 0.5 points (89.9% and 90.2%) despite one teacher having 7 times more parameters than the other, while self-distillation reaches 90.4% (Table 10). All three students remain $\sim$8 points below the no-KD WS-206K baseline of 98.0%, so teacher quality is not the binding constraint on the KD degradation. Results are single-seed.

Table 10: Teacher quality (student is WS-206K, 1 seed).

Teacher	Params	Student validity (%)	FCD
GPT-3.2M	3.2M	89.9	2.18
GPT-467K-mod	467K	90.2	2.26
WS-206K (self)	206K	90.4	2.25

I Direct Preference Optimization

We apply DPO to two weight-shared bases (WS-206K, WS-82K) and one curriculum-distilled WS-206K variant, using the settings in Appendix A. Across all three bases, validity changes by less than 2 percentage points and FCD shifts by at most 0.16 (Table 11). All results are single-seed.

Table 11: DPO before/after results (1 seed each).

Base	Validity pre (%)	Validity post (%)	FCD pre	FCD post
WS-206K	98.5	97.6	2.52	2.58
WS-82K	96.7	96.5	2.58	2.74
WS-206K (curriculum-distilled)	91.0	92.6	2.29	2.42