ATTENTION TRACKS REACTION CENTERS, BUT DOESN'T NEED THEM: A CAUSAL STUDY OF A RETROSYNTHESIS TRANSFORMER

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention weights in transformer-based chemistry models are widely treated as an interpretability signal — attention concentrated on a molecule's reaction center is commonly cited as evidence that a model has "learned" the relevant chemistry. We test this assumption directly using causal intervention. In a retrosynthesis Chemformer trained on USPTO-50k, encoder attention correlates strongly with ground-truth reaction centers (within-molecule AUC 0.902 vs. 0.836 for hand-engineered structural descriptors), replicating the correlational finding that motivates this interpretability practice. However, zero-ablating attention heads and measuring the resulting change in generation likelihood reveals that causal necessity is inversely related to this correlation: the heads most predictive of reaction centers are among the least necessary for generation, while 74.9% of total ablation damage concentrates in the encoder's first layer — the layer with the lowest reaction-center correlation of any layer in the network. We show this necessity is not decomposable into individual head contributions: the sum of single-head ablation effects in layer 1 is 115$\times$ smaller than the effect of ablating all eight heads jointly, and which single head survives a near-total ablation determines whether the model recovers 84% or 31% of baseline performance. We trace this structure to a measurable property of each head: attention entropy predicts sole-survivor sufficiency ($\rho = -0.857$) in the opposite direction from necessity, explaining why sharp, low-entropy positional heads are individually redundant yet individually sufficient, while diffuse, high-entropy heads are collectively necessary yet cannot reconstruct function alone. Our results extend the attention-is-not-explanation literature from NLP classification into generative chemistry models, and show that correlational interpretability claims in this domain — as in concurrent findings on 3D-geometry alignment in other chemical language models — do not reliably predict where causal computation actually occurs.

1 INTRODUCTION

Attention weights are among the most visible interpretability artifacts a transformer produces, and in chemistry-focused models they are used more literally than almost anywhere else: attention heatmaps overlaid on a molecule are routinely presented as showing which atoms the model considers chemically important, with high attention on a reaction center offered as evidence a reaction-prediction model has learned the underlying chemistry (Schwaller et al., 2019; 2021). This practice inherits an assumption already contested once — in NLP, Jain & Wallace (2019) showed attention weights are frequently uncorrelated with other measures of feature importance and that models can produce very different attention distributions while making identical predictions, prompting a still-unresolved methodological debate (Wiegreffe & Pinter, 2019; Serrano & Smith, 2019) fought almost entirely with correlational and perturbation-based diagnostics on NLP classification models. Whether its central lesson — that attention correlating with something meaningful does not imply causal responsibility for that meaning — transfers to a generative chemistry model reading SMILES strings has not been directly tested.

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.

We test it directly, using causal ablation rather than correlational diagnostics, on a retrosynthesis Chemformer (Irwin et al., 2022) fine-tuned on USPTO-50k. We first confirm the correlational premise the field relies on: encoder self-attention predicts ground-truth reaction centers substantially better than hand-engineered structural descriptors (AUC 0.902 vs. 0.836). We then ask the causal question directly: does zero-ablating a reaction-center-correlated head actually damage the model's ability to predict reactants? The answer is inverted. The heads most correlated with reaction centers are among the least costly to remove, while 74.9% of the effect of ablating every head in the network concentrates in encoder layer 1 — the layer with the lowest reaction-center correlation of any we measured.

This dissociation raises an immediate follow-up: what is layer 1 doing, if not tracking chemically meaningful structure? Its causal dominance is not a simple sum of independently important heads — the eight single-head effects in layer 1 sum to roughly 1/115th of the damage from ablating all eight jointly, and a 7-of-8 "leave-one-in" sweep shows the surviving head's identity determines 84% vs. as little as 31% function retention, a 4.2$\times$ spread ruling out a redundant-ensemble picture. We trace this to attention entropy: low-entropy, positional heads are individually sufficient as sole survivors but collectively redundant, while high-entropy, diffuse heads are collectively necessary yet individually insufficient ($\rho = -0.857$ across the eight layer-1 heads) — to our knowledge a previously uncharacterized mechanism linking an observable attention statistic to causal structure.

These results extend the attention-is-not-explanation literature (Jain & Wallace, 2019) into generative chemistry models, and connect to concurrent evidence that a separate correlational claim — MOLFORMER's attention tracking 3D molecular geometry — similarly fails causal analysis (Hao et al., 2026), suggesting this gap is not an artifact of one model or analysis choice.

Our contributions are:

- The first causal (rather than correlational) test of treating chemistry-transformer attention as evidence of learned reaction chemistry, via zero-ablation on a fine-tuned Chemformer.
- Evidence that correlational salience and causal necessity are inversely related at the head level, with necessity concentrated in the layer of lowest salience.
- A characterization of this necessity as a non-additive, threshold-like property of layer 1 as a whole, not any single bottleneck head.
- Attention entropy as a mechanistic explanation for the dissociation, predicting head sufficiency and necessity in opposite directions.
- Several negative results reported with full effect sizes and CIs — a null test of class-specific specialization and a null contribution from electronic (Gasteiger) descriptors.

2 Related Work

Attention as explanation. Whether attention weights constitute a faithful account of model behavior has been contested since Jain & Wallace (2019) showed that learned attention distributions in NLP classifiers are often uncorrelated with gradient-based importance measures and that very different attention patterns can produce equivalent predictions, prompting pushback on the definition of explanation used (Wiegreffe & Pinter, 2019) and a related critique via attention erasure (Serrano & Smith, 2019). This debate was conducted almost entirely through correlational and perturbation-based diagnostics on recurrent and early transformer NLP models. We extend it using direct causal intervention (zero-ablation with teacher-forced likelihood) on a structurally different input domain — SMILES-encoded molecules.

Head-level ablation and pruning. A parallel literature established that transformer attention heads contribute unequally to task performance: Michel et al. (2019) showed a large fraction of heads in trained machine translation and BERT models can be removed with minimal performance loss, and Voita et al. (2019) reached a complementary conclusion via learned pruning, finding a small number of specialized heads do most of the work. Neither directly contrasts causal necessity against correlational salience for the same heads, nor examines why some heads are individually replaceable while others are not. Our leave-one-in and cumulative-ablation results speak directly to this gap, showing a 115$\times$ non-additive gap between individual and group head effects in an early encoder

layer — a structure not characterized in a literature that, by construction, studies heads largely in isolation.

Causal methods in mechanistic interpretability, and chemistry attention. Causal intervention is a central tool for localizing computation in transformers broadly; Meng et al. (2022) use corrupt-and-restore causal tracing to localize factual recall to mid-layer feed-forward modules in autoregressive LLMs, and our zero-ablation methodology sits in the same tradition, applied instead to encoder self-attention in a sequence-to-sequence chemistry model. Within cheminformatics specifically, attention has been used as an interpretability signal since Schwaller et al.'s Molecular Transformer (Schwaller et al., 2019; 2021), which reported that attention qualitatively tracks reaction centers — a claim subsequent retrosynthesis work has largely repeated uncritically. This is precisely the claim our results complicate: the reaction-center-correlated heads in Chemformer (Irwin et al., 2022) are among the least causally necessary, while the causally dominant heads concentrate in the layer least correlated with reaction centers of any we tested.

Concurrent work. Closest to our work, Hao et al. (2026) apply a mechanistic lens to MOLFORMER (Ross et al., 2022) and SMI-TED (Soares et al., 2024), similarly finding that a widely-cited correlational claim — that MOLFORMER's attention tracks 3D molecular geometry — fails under careful analysis, using linear probing and causal steering to show linear decodability and causal usability diverge by depth. Our work differs in task (retrosynthesis vs. masked-LM property encoding), intervention (head-level zero-ablation on generation likelihood vs. representation-level steering), and target (reaction centers vs. 3D geometry); where they show causal usability lags decodability across depth, we show a within-layer finding — necessity is non-decomposable and anti-correlated with salience at the individual-head level — that probing is not positioned to surface. Together, the two results suggest correlational interpretability signals in chemical language models are systematically weaker guides to internal computation than the field's attention-visualization era assumed.

3 METHODS

3.1 DATASET AND REACTION CENTER EXTRACTION

We use USPTO-50k in the retrosynthesis direction (encoder reads the product SMILES; decoder generates reactants). Reaction centers are bonds present in the product but absent, or changed in order, between the same atom-mapped atoms in the reactants; hydrogen-count or formal-charge changes are excluded from this label, since these typically reflect unmapped leaving groups (e.g., deprotonation, halide loss) rather than intramolecular bond reorganization. Of 38,782 mapped reactions, 10,854 lack a reaction center under this definition — predominantly fully-unmapped leaving groups — and are excluded. The retained 27,928 reactions have a mean of 2.13 reaction centers per product (9.2% positive-class rate), using the dataset's standard train/val/test split (31,031/3,862/3,889).

3.2 MODEL AND ATTENTION EXTRACTION

We fine-tune Chemformer (Irwin et al., 2022) (BART-style, 6 encoder layers, 8 heads/layer, 64-dimensional head projections) on USPTO-50k retrosynthesis, selecting the minimum-validation-loss checkpoint. Encoder self-attention is extracted via forward hooks on each layer's `self_attn` module, aggregated as mean incoming attention per atom token across all query positions. All 38,782 reactions are processed with zero token-atom alignment failures.

3.3 CORRELATIONAL ANALYSIS

We evaluate attention's prediction of reaction-center membership via within-molecule AUC — a per-molecule Mann-Whitney statistic averaged over the 3,889 test molecules, avoiding the cross-molecule confound of a pooled atom-level AUC — across five nested feature sets: positional (token index), structural (RDKit graph descriptors), structural + Gasteiger partial charges, attention alone, and attention + structural, plus a shuffled-attention control for the chance floor. We also compute per-head Spearman correlations to reaction-center membership and node degree, to flag degree-confounded heads.

3.4 CAUSAL ABLATION

We ablate a head by zeroing the corresponding 64-column block of the encoder self-attention output projection (`self_attn.out_proj.weight`) — verified equivalent to zeroing that head's output post-hoc, since PyTorch's `MultiheadAttention` invokes `out_proj` functionally rather than as an interceptable module call. Our damage metric is the paired change in mean per-token teacher-forced log-likelihood (product $\rightarrow$ reactants) on 2,000 held-out molecules pooled from validation and test (seed 42), validated with sanity checks (determinism under no ablation, exact baseline restoration on clearing the mask, uniformly large damage from ablating all 48 heads). We report four experiment classes: full-layer ablation (localizing damage by depth); within-layer single-head and cumulative ablation (testing additivity); "leave-one-in" ablation (7-of-8 heads removed, testing interchangeability); and correlational-rank head-set ablation (top/bottom-5 by reaction-center or degree correlation, plus matched random sets, contrasting salience against necessity). We report the full paired distribution across molecules, not only the mean, since roughly 40% of molecules improve in log-likelihood under head-level ablation.

3.5 ATTENTION CHARACTERIZATION AND STATISTICAL REPORTING

On a separate 400-molecule held-out sample, we characterize each head's attention via offset profile (positional concentration), row entropy (normalized to 0–1, via forward hooks with `need_weights=True`, `average_attn_weights=False`), and token-class enrichment (atom, ring-closure, branch, bond, special, or fragment tokens, relative to base-rate frequency). We test, via Spearman correlation over the 8 layer-1 heads, whether entropy, positional concentration, or token-class enrichment predicts single-head ablation damage (necessity) or leave-one-in sole-survivor performance (sufficiency). Given the scale of the correlational tests ($n \approx 31{,}000$), standard significance testing is uninformative (nearly all comparisons reach significance after FDR correction), so we report effect sizes with 95% confidence intervals throughout, reserving p-values for the small-n ($n = 8$) entropy analyses, where they retain interpretive value but support only a suggestive rather than confirmatory claim.

4 RESULTS

4.1 ATTENTION TRACKS REACTION CENTERS, BUT WITH A CONFOUND

We first establish the correlational premise that motivates attention-based interpretability in chemistry models. Across five nested feature sets, encoder attention substantially outperforms hand-engineered structural descriptors at predicting reaction-center membership on a within-molecule basis: positional features alone reach AUC 0.622, structural descriptors 0.836, structural descriptors plus Gasteiger partial charges 0.837, attention alone 0.902, and attention combined with structural features 0.914 (Figure 1). The gain of attention over structural descriptors alone is 0.077 (95% CI [0.075, 0.080]), scored across 3,889 test molecules. A shuffled-attention control — attention values randomly permuted within each molecule — falls to 0.502, at chance, confirming the signal is not an artifact of molecule-level covariates. Gasteiger partial charges add essentially nothing beyond structural descriptors (+0.001), a first indication that electronic structure, at least as approximated by this simple charge model, is not what the network's attention is tracking.

Breaking this correlation down by individual head (Figure 6, Appendix) reveals a two-population structure. The single head most correlated with reaction-center membership, L5H8, reaches $\rho = 0.282$, but also correlates with node degree ($\rho = 0.055$), a potential confound since reaction centers in USPTO-50k are not degree-uniform. A cleaner signal comes from L5H6, which shows the strongest correlation of any head to a different target — atom-to-reaction-center distance ($\rho = -0.446$, solo AUC 0.825) — suggesting the network encodes not just "is this a reaction center" but graded proximity to one. We flag this L5-heavy concentration of correlational salience explicitly because it becomes the direct counterexample in Section 4.2: these are precisely the heads the causal ablation results identify as among the least necessary.

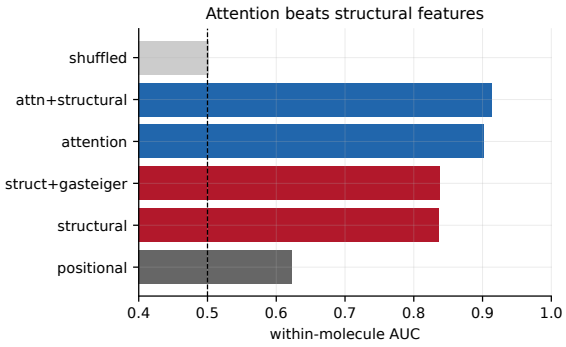

Figure 1: Within-molecule AUC for predicting reaction-center membership across nested feature sets, including a shuffled-attention control. Attention substantially outperforms hand-engineered structural descriptors, and combining the two adds a further, smaller gain.

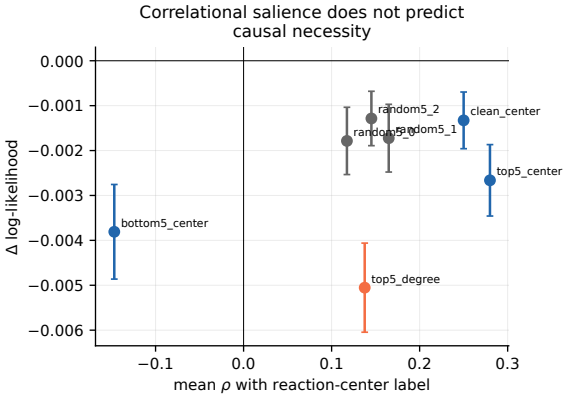

Figure 2: Head-set mean reaction-center correlation vs. causal damage (paired log-likelihood change), with 95% confidence intervals. The ordering inverts: the most correlated head set is the least causally necessary.

4.2 Correlational salience does not predict causal necessity — and the relationship inverts

We next test whether the heads most correlated with reaction centers are also causally responsible for the model's retrosynthesis predictions. Ablating the five heads most correlated with reaction centers ("top5_center," mean $\rho = 0.280$) produces a mean log-likelihood change of -0.00266. Ablating the five heads most correlated with node degree instead ("top5_degree," mean $\rho = 0.138$ — a substantially weaker correlational signal) produces roughly twice the damage: -0.00505. The most reaction-center-correlated head set we tested, "clean_center" ($\rho = 0.250$, selected to exclude degree-confounded heads), is the least causally necessary of every head set evaluated (-0.00133) — meaning the purest correlational signal in the entire network corresponds to the least causally load-bearing computation (Figure 2).

This inversion holds at the level of individual heads as well as head sets. L5H8, the single head with the highest reaction-center correlation ($\rho = 0.282$), causes only -0.00058 log-likelihood damage when ablated alone. L6H1, a head with modest correlational salience, improves the model's output when removed (delta $= +0.00026$) — a small but genuine negative-necessity effect, not merely noise around zero. A head set selected for low correlation to reaction centers ("low_rho5," three of whose five heads happen to fall in layer 1) causes the largest single-head-set damage we measured among the four-head comparisons, -0.00991 — roughly 4–8$\times$ the damage of the high-correlation sets, foreshadowing the layer-1 result in Section 4.3.

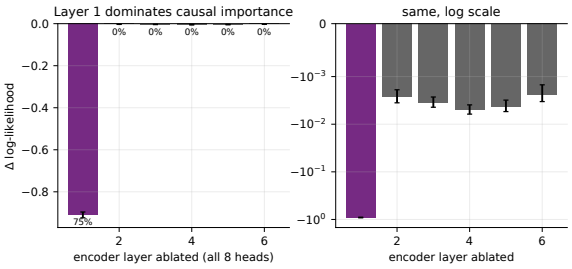

Figure 3: Log-likelihood change from ablating each encoder layer in isolation (all 8 heads), linear and symlog scale. Layer 1 accounts for 74.9% of the damage from ablating all 48 heads.

We treat this as a genuine dissociation rather than a fragile artifact of small-n comparisons: the ordering top5_degree $>$ clean_center is consistent across every related head-set comparison we ran, and Figure 12 (Appendix) shows the fraction of molecules degraded (rather than improved) under ablation clusters around 0.55–0.65 for essentially every head-level condition tested — the inversion is not being driven by a handful of molecules with outsized log-likelihood swings.

4.3 Necessity concentrates in layer 1 — precisely where correlational salience is lowest

Ablating an entire encoder layer (all 8 heads) reveals a stark concentration of causal damage. Layer 1 alone accounts for -0.914 of the -1.220 log-likelihood change caused by ablating all 48 heads in the network — 74.9% of total damage from one-sixth of the heads. Layers 2 through 6 combined (40 heads) account for the remaining 38% (-0.465; the two contributions do not sum to exactly 100% due to interaction effects between layers, discussed below). Individually, layers 2 through 6 each cause between -0.002 and -0.004 log-likelihood damage — two orders of magnitude below layer 1 (Figure 3).

This concentration is the direct inverse of the correlational picture in Section 4.1: layer 1 has the lowest mean reaction-center correlation ($\rho_{\text{center}} = 0.074$) of any of the six encoder layers, while the correlationally richest layer, layer 5 (home to L5H6 and L5H8), causes comparatively little damage when ablated as a whole. Maximum causal necessity co-locates with minimum correlational salience, layer for layer, not just head for head.

A further non-monotonicity emerges from cumulative ablation: ablating layers 1–3 together (-1.384) produces *more* damage than ablating all 48 heads across all 6 layers (-1.220), implying later layers partially repair the corruption from ablating layers 1–3 rather than adding independent damage. A marginal-damage analysis confirms this: given layer 1 already ablated, each additional layer's marginal cost roughly halves with depth (layer 2: -0.381; layers 3 through 6: -0.132, -0.056, -0.022, -0.009). The model's downstream layers are not independent damage-accumulators; they appear to actively compensate for a corrupted early representation, and removing that compensatory machinery alongside the original corruption is, counterintuitively, less damaging than removing the corruption alone.

4.4 Layer 1 is a non-additive threshold circuit, not a redundant ensemble

Given layer 1's outsized causal role, we ask whether it functions as eight independently important heads or as a more tightly coupled unit. The evidence strongly favors the latter. Summing the eight single-head ablation effects in layer 1 yields -0.0079; ablating all eight jointly yields -0.914 — a 115$\times$ gap between the sum of parts and the whole (Figure 5, Appendix). Cumulative ablation in order of decreasing individual effect shows this gap opens sharply rather than gradually: ablating the seven least-damaging-individually heads together costs only -0.215, but adding the eighth (L1H1) jumps the cumulative damage to -0.914, a 4.3$\times$ single-step increase. This is a threshold, not a slope — the layer's function degrades gracefully until nearly all heads are gone, then collapses.

The "leave-one-in" experiment (ablate 7 of 8 heads, vary which one survives; Figure 7, Appendix) shows these heads are also not interchangeable substitutes for one another. Recovery ranges from

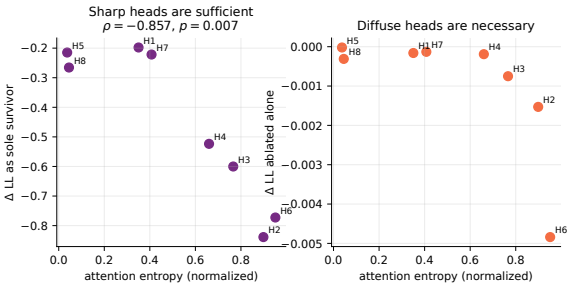

Figure 4: Attention entropy vs. sole-survivor damage (sufficiency) and vs. ablate-alone damage (necessity) for the 8 layer-1 heads. Entropy predicts the two in opposite directions ($\rho = -0.857$ for sufficiency).

84% of baseline function when L1H1 alone survives (delta -0.198, versus -1.220 for full ablation) to just 31% when L1H2 alone survives (delta -0.839) — a 4.2× spread depending purely on which single head is spared. As a control, we repeat the leave-one-in procedure on layer 4, a layer with comparatively little causal weight: results are uniformly small (-0.003 to -0.004) regardless of which head survives, confirming the all-or-nothing, non-interchangeable structure in layer 1 is a specific property of that layer, not a generic artifact of gutting any 7-of-8 head set.

Notably, L1H6 — the single head causing the most damage when ablated alone (-0.0048, roughly 60% of the -0.0079 sum of all eight singles) — is not a bottleneck in the leave-one-in sense: as sole survivor it recovers only 21% of function, and it accounts for only 0.5% of the full group's joint effect. The head most damaging to remove individually and the head most valuable to keep as a sole survivor are not the same head.

4.5 Mechanism: attention entropy separates necessity from sufficiency

Sections 4.3–4.4 establish where necessity concentrates and that it is non-additive; we now identify what distinguishes individually-sufficient heads (good sole survivors) from individually-necessary-but-non-substitutable heads. Characterizing each layer-1 head's attention pattern on a held-out sample of 400 molecules reveals that four of the eight are near-deterministic positional shift operators (e.g., L1H5 places 93.9% of its mass one token to the left, row entropy below 0.05), while the remainder are comparatively diffuse, with entropy up to 0.90 (Figure 8, Appendix).

This entropy measurement predicts sole-survivor sufficiency almost perfectly in the anti-correlated direction we would not have guessed from the necessity data alone: Spearman $\rho = -0.857$ ($p = 0.007$, $n = 8$ heads) between a head's attention entropy and its damage when left as sole survivor (Figure 4). L1H1, the best sole survivor (84% recovery), has entropy 0.351 — sharp and positional. L1H2, the worst sole survivor (31% recovery), has entropy 0.899 — nearly uniform. The same relationship holds in reverse for single-head necessity: L1H6, the most damaging head to ablate alone, has entropy 0.951, among the highest in the layer.

This gives the necessity/sufficiency dissociation a mechanistic explanation rather than leaving it as an unexplained structural curiosity. Low-entropy, positional heads are collectively redundant — several compute similar local shift operations, so any one is individually dispensable — but that same sharpness makes each one, alone, capable of approximately reconstructing what the layer needs. High-entropy, diffuse heads are collectively indispensable — they appear to aggregate information no positional head captures — but no single diffuse head, alone, has a sharp enough function to substitute for the layer's role.

We separately tested whether what a head attends to — rather than how it attends — predicts causal role, by correlating each head's token-class enrichment with necessity and sufficiency. All four correlations were null ($|\rho| \leq 0.405$, all $p > 0.3$), including for ring-closure enrichment specifically ($\rho = -0.262$, $p = 0.531$) — the class of token we might expect a priori to be syntax-critical for maintaining valid SMILES ring-bond matching. Only the distributional shape of attention, not its semantic target, predicts causal role here.

Consistent with the broader depth gradient reported in Section 4.3, local positional attention (offset ± 1 mass) is strongest in layers 1–2 and decays roughly threefold by layer 6, while attention to special tokens (BOS/EOS) rises roughly fourfold over the same range (Figure 13, Appendix) — early layers appear to specialize in local sequential structure, with later layers pooling into sequence-level slots.

4.6 Negative results: class specialization and electronic structure

Two correlational patterns that might plausibly translate into causal specialization do not. First, as noted in Section 4.1, Gasteiger partial charges contribute no measurable improvement over structural descriptors alone in predicting reaction centers (+0.001 AUC). Second, several heads show correlational specialization to specific USPTO-50k reaction classes in prior probing analysis; we test whether this specialization is causal by comparing ablation damage on held-out molecules from the specialized class against damage on other classes. All four tested contrasts' 95% CIs span zero (L6H1: diff -0.000129; L2H6: diff -0.00018; L5H6: diff -0.00033; top-5 degree-correlated set: diff $+0.001616$, $p = 0.0028$, the largest point estimate but still CI-crossing zero). Removing the 20 heads most correlated with a single reaction class (42% of the entire network) costs only 6.1% of the damage caused by ablating all 48 heads — far below what class-specific circuitry of that scale would predict if the correlational specialization reported in prior probing work were causally load-bearing. We report this as a clean negative result: correlational class specialization does not imply causal class specialization at the level of attention heads.

5 Discussion

5.1 What we found, and what we did not find

The central result of this paper is a dissociation: the attention pattern most correlated with a chemically meaningful ground truth (reaction centers) is among the least causally responsible for the model's predictions, while the causally dominant computation sits in the layer least correlated with that ground truth. This is not a claim that attention is uninformative — Section 4.1 shows attention substantially outperforms hand-engineered structural descriptors at predicting reaction centers, a real correlational signal — but a claim that this signal is a poor guide to where computation actually happens, conflating two questions our causal results show can point in opposite directions. It is also worth being precise about what Section 4.5's entropy mechanism does not explain: it predicts which heads are individually sufficient versus necessary, but not why layer 1's contribution is chemically relevant in any semantic sense. We have a mechanistic account of the layer's internal structure without a mechanistic account of what chemical computation it performs; we view closing that gap as future work, not a claim the current results already make.

5.2 Two overreaches we do not make

Layer 1 is not a syntax parser: Figure 8's positional heads attend almost deterministically to the immediately adjacent token — a local sequential-context circuit, not evidence of tracking the long-range dependencies (ring-closure matching, parenthesis balancing) formal parsing requires — and Section 4.5's null result on token-class enrichment confirms attending preferentially to ring-closure tokens does not predict causal role. What we can defensibly claim is narrower: layer 1 builds local sequential context via a few near-deterministic positional operators, disproportionately load-bearing downstream. Nor are its heads redundant in the sense pruning literature would assume: Section 4.4's near-zero single-head ablation effects might suggest individual heads are safely prunable, but the leave-one-in results contradict this — heads are redundant only in that removing any one barely matters while seven others compensate, not interchangeable, since which single head survives determines a $4.2\times$ spread in recovered function. A pruning method relying on single-head ablation scores alone (Michel et al., 2019; Voita et al., 2019) would rank all eight as equally low-priority, missing that some "unimportant" heads are the layer's only path to recovery under aggressive pruning.

5.3 Scope, convergent evidence, and practical implications

Every result here concerns encoder self-attention, not decoder self- or cross-attention, both mechanistically active during generation — a meaningful limitation, since cross-attention is architecturally the more direct route by which the decoder reads the product representation, and closing this gap

(Section 6) is a priority. Our results also converge with concurrent, independent findings: Hao et al. (2026) show a different correlational claim (MOLFORMER attention tracking 3D geometry) similarly fails causal scrutiny, via a different model, intervention, and target — less a claim of a universal law than a signal that chemistry's attention-visualization-era practices were adopted without the causal validation step NLP interpretability already underwent. Practically: an attention heatmap lighting up on a reaction center is evidence the model learned a correlate of reaction centers, not evidence of which part of the network is responsible for its predictions, and should not serve as a debugging or trust signal without independent causal validation.

6 Future Work

Several limitations point to natural extensions. **Scope:** our results concern encoder self-attention only; decoder self- and cross-attention were not characterized, and cross-attention in particular is the more natural place to expect causal reaction-center tracking, since it is architecturally how the decoder reads the product representation during generation. **Method:** we use zero-ablation throughout (Michel et al., 2019; Voita et al., 2019), which is standard but pushes the model off-distribution; our qualitative findings (layer-1 dominance, the threshold structure, the necessity/sufficiency dissociation) should be checked against a distribution-respecting ablation such as mean- or resample-ablation before being treated as method-independent. **Generalization:** all results come from a single fine-tuned Chemformer checkpoint, so it is unknown whether the layer-1 circuit and entropy mechanism are properties of this model, of BART-style chemistry transformers generally, or of SMILES sequence-to-sequence architectures broadly — future work should repeat this pipeline across architectural variants (positional encoding, pretraining objective, tokenization) and seeds. **Power:** the entropy–sufficiency correlation ($\rho = -0.857$) is computed over only $n = 8$ heads in one layer, and our null result on class-specific specialization (Section 4.6) used a coarse reaction taxonomy; both would benefit from more heads, layers, and models. We deliberately do not attempt to package the entropy–sufficiency relationship into a general ablation-free importance proxy here (Voita et al., 2019), since doing so responsibly requires the cross-architecture, higher-power evidence above.

7 Conclusion

We tested a correlational interpretability practice common in chemistry transformers — that attention concentrated on a molecule's reaction center indicates the model has learned that chemistry — against direct causal intervention, in a retrosynthesis Chemformer fine-tuned on USPTO-50k. The two measures disagree, and disagree in a specific and informative direction: the heads most correlated with reaction centers are among the least causally necessary for generation, while 74.9% of total causal damage concentrates in the one encoder layer with the lowest reaction-center correlation of any layer in the network. This is not a claim that attention carries no chemical signal — it does, and outperforms hand-engineered structural descriptors at predicting reaction centers — but a claim that this signal is a poor guide to where the model's computation is actually load-bearing.

Beyond the dissociation itself, the causally dominant layer is not a simple aggregate of independently important heads but a threshold circuit: single-head effects sum to roughly 1/115th of the group effect, and which head survives near-total ablation determines a 4.2× spread in recovered function, traced to attention entropy predicting sufficiency and necessity in opposite directions. Two patterns that might plausibly have been causal — electronic-descriptor contribution and class-specific specialization — were not, and we report both as clean negative results rather than omitting them. Every finding concerns encoder self-attention specifically; whether these dynamics hold in decoder attention, under distribution-respecting ablation, or across chemistry-transformer architectures more broadly are open questions (Section 6). Together with concurrent evidence that a different correlational claim fails similarly in a different chemical language model (Hao et al., 2026), our results suggest this gap is not a peculiarity of one model, and that correlational interpretability claims in this domain warrant the same causal scrutiny NLP's own attention-as-explanation debate eventually applied to itself.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting and editing manuscript text based on results, figures, and provenance records produced by the authors' own experiments, and to assist with LaTeX formatting of the submission. We have not used generative AI tools to design experiments, generate or fabricate data, or produce the underlying figures, all of which were produced directly from the authors' own model training, ablation, and analysis code. All AI-assisted text was reviewed, fact-checked against the underlying experimental provenance records, and edited by the authors prior to submission. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Full experimental details — dataset construction, ablation methodology, and statistical procedures — are provided in Section 3; hyperparameters and checkpoint selection criteria are described in Section 3 and Appendix figures are generated directly from the same ablation and probing code described there.

REFERENCES

Wanqing Hao, Zetong Xu, Zhiyuan Han, Xiaoyan Bai, and Chenhao Tan. Mechanistic interpretability of chemical language models: Molecular geometry and representations in MoLM-Former and SMI-TED. In *ICML 2026 Workshop on Mechanistic Interpretability*, 2026. URL https://openreview.net/forum?id=iuU13T2BAz.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

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 (NAACL-HLT)*, pp. 3543–3556, 2019.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3:144–152, 2021.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2931–2951, 2019.

Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, Kristin Schmidt, and Renato Vasconcelos. A large encoder-decoder family of foundation models for chemical language. *arXiv preprint arXiv:2407.20267*, 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5797–5808, 2019.

Sarah Wiegreffe and Yuval Pinter. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

A Appendix

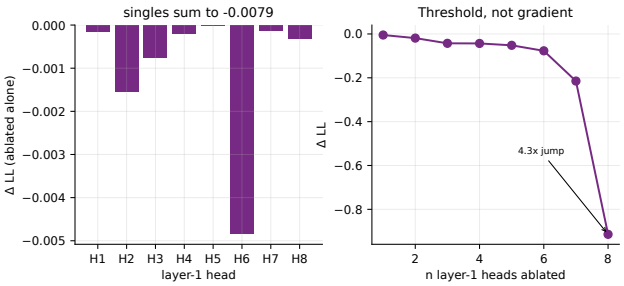

Figure 5: Layer-1 single-head ablation effects summed (left) vs. cumulative ablation in order of decreasing individual effect (right). The sum of parts is 115× smaller than the joint group effect. Referenced in Section 4.4.

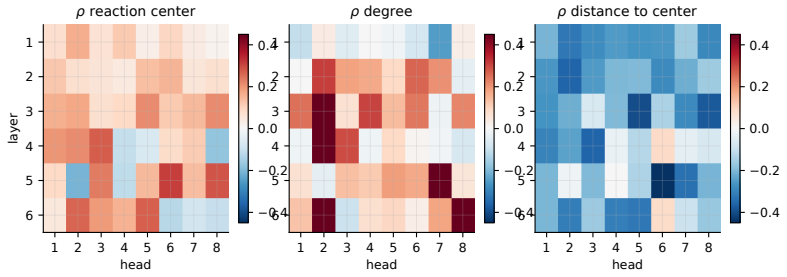

Figure 6: Per-head Spearman correlation heatmaps (reaction-center membership, node degree, and reaction-center distance) across all 6 layers × 8 heads. Correlational salience concentrates in later layers, especially layer 5. Referenced in Section 4.1.

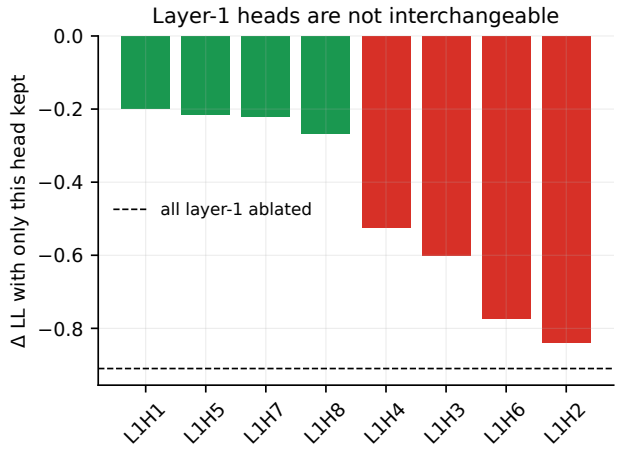

Figure 7: "Leave-one-in" ablation: log-likelihood damage as a function of which single head survives a 7-of-8 layer-1 ablation, versus a flat layer-4 control. Referenced in Section 4.4.

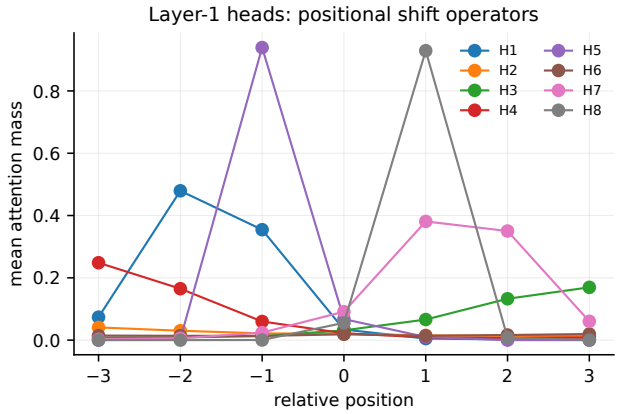

Figure 8: Per-head offset profiles for all 8 layer-1 heads. Four heads are near-deterministic positional shift operators; the remainder are diffuse. Referenced in Section 4.5.

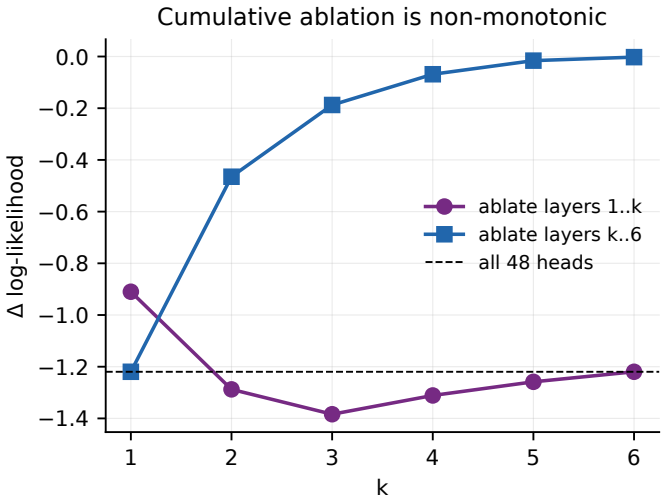

Figure 9: Bottom-up (layers $1..k$) vs. top-down (layers $k..6$) cumulative ablation, illustrating the non-monotonicity discussed in Section 4.3: ablating layers 1–3 causes more damage than ablating all 6 layers.

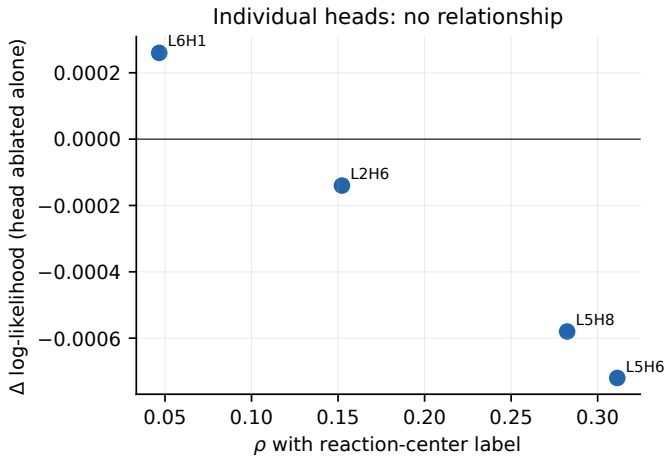

Figure 10: Four named heads: reaction-center correlation vs. single-head ablation damage, showing the dissociation holds at the individual-head level.

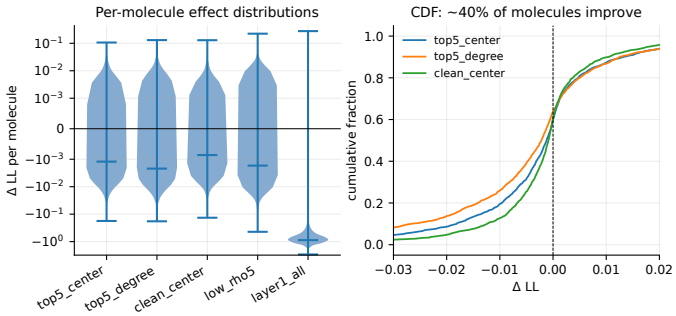

Figure 11: Violin plots and CDFs of per-molecule log-likelihood change across ablation conditions, showing that roughly 40% of molecules improve under head-level ablation.

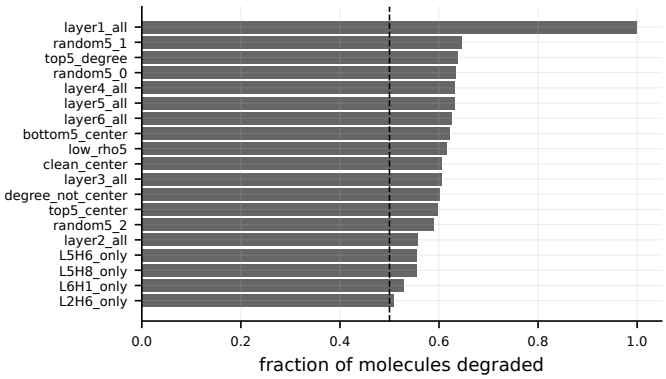

Figure 12: Fraction of molecules degraded (log-likelihood decreased) across all ablation conditions tested, sorted.

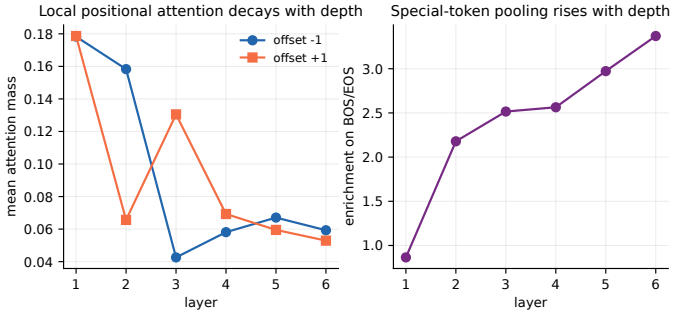

Figure 13: Offset-1/+1 attention mass and BOS/EOS enrichment by layer, showing local positional attention decaying and special-token pooling rising with depth.