Mechanistic Interpretability of Chemical Language Models: Molecular Geometry and Representations in MoLFormer and SMI-TED

Wanqing Hao [1] **Zetong Xu** [1] **Zhiyuan Han** [1] **Xiaoyan Bai** [1] **Chenhao Tan** [1]

Abstract

Chemical language models trained on SMILES strings, text-based representations of molecules, achieve strong performance on chemical prediction tasks, yet it remains unclear how they internally represent molecular structure and chemical properties. In this work, we analyze two widely used models, MoLFormer (Ross et al., 2022) and SMI-TED (Soares et al., 2024), from a mechanistic interpretability perspective. Prior analyses of chemical foundation models have largely focused on correlational methods, including attention correlations, embedding structure, and input-output associations. We take a step toward a more mechanistic understanding by combining layerwise attention analysis, linear probing, and causal intervention. We find that attention exhibits only weak and selective alignment with underlying molecular geometry. Probing further reveals that SMI-TED encodes contextual chemical properties as cleaner linear directions than MoLFormer. Moving beyond correlational analysis, our causal interventions show that these directions are functionally used during computation, and earlier in SMI-TED than in MoLFormer. Despite strong downstream performance, our results suggest that successful SMILES pretraining does not necessarily produce spatially grounded or uniformly usable chemical representations, highlighting how diverse internal representations can support similar predictive success.

1. Introduction

Foundation models pretrained on large scientific datasets are rapidly transforming prediction across the natural sciences. For example, they have achieved promising results in domains including proteins (Lin et al., 2023a; Jumper et al., 2021), genomics (Nguyen et al., 2024), and weather forecasting (Lam et al., 2023). Specifically, chemistry foundation models pretrained on large-scale SMILES corpora, including ChemBERTa (Chithrananda et al., 2020), MoLFormer (Ross et al., 2022), SMI-TED (Soares et al., 2024), and ChemFM (Cai et al., 2026), now serve as standard backbones for molecular property prediction, ADMET screening, and materials discovery (Wu et al., 2018; Harod et al., 2025; Sharma et al., 2025). Despite operating directly on molecular text without explicit geometric supervision, these models achieve strong performance across a wide range of downstream chemistry tasks.

However, strong downstream performance does not imply that we understand what these models have learned internally. SMILES strings do not explicitly encode many chemically meaningful properties. For example, atom identity and aromaticity are directly visible from the token sequence, but higher-level structure, such as stereochemistry, valence, ring structure, and 3D geometry, must instead be inferred from context. Thus, it is unclear whether strong downstream performance reflects meaningful chemical representations or weaker statistical patterns learned during pretraining. Because molecular models are typically deployed as black-box feature extractors, three concrete questions about their internal mechanism remain open: *(i)* do attention heads track 3D molecular structure? *(ii)* at what depth do contextual chemical properties such as ring membership or chirality emerge in the residual stream? and *(iii)* are these representations actively used during computation, or merely co-encoded with the eventual prediction?

To answer these questions, we apply the same set of analyses to two of the most widely used SMILES encoders, MoLFormer and SMI-TED. The models are architecturally similar as both of them use 12-layer transformer encoders with 768-dimensional hidden states and rotary linear attention. This makes them a useful comparison pair for isolating how different pretraining designs shape internal chemical representations. Across both models, we perform three complementary experiments: (1) attention–distance correlation across all heads to test whether attention reflects 3D molecular geometry; (2) linear probing of eight per-atom RDKit-derived chemical properties (Landrum et al.,

[1]University of Chicago. Correspondence to: Xiaoyan Bai <smallyan@uchicago.edu>, Wanqing Hao <anitahao@uchicago.edu>.

Accepted at Mechanistic Interpretability Workshop at ICML 2026

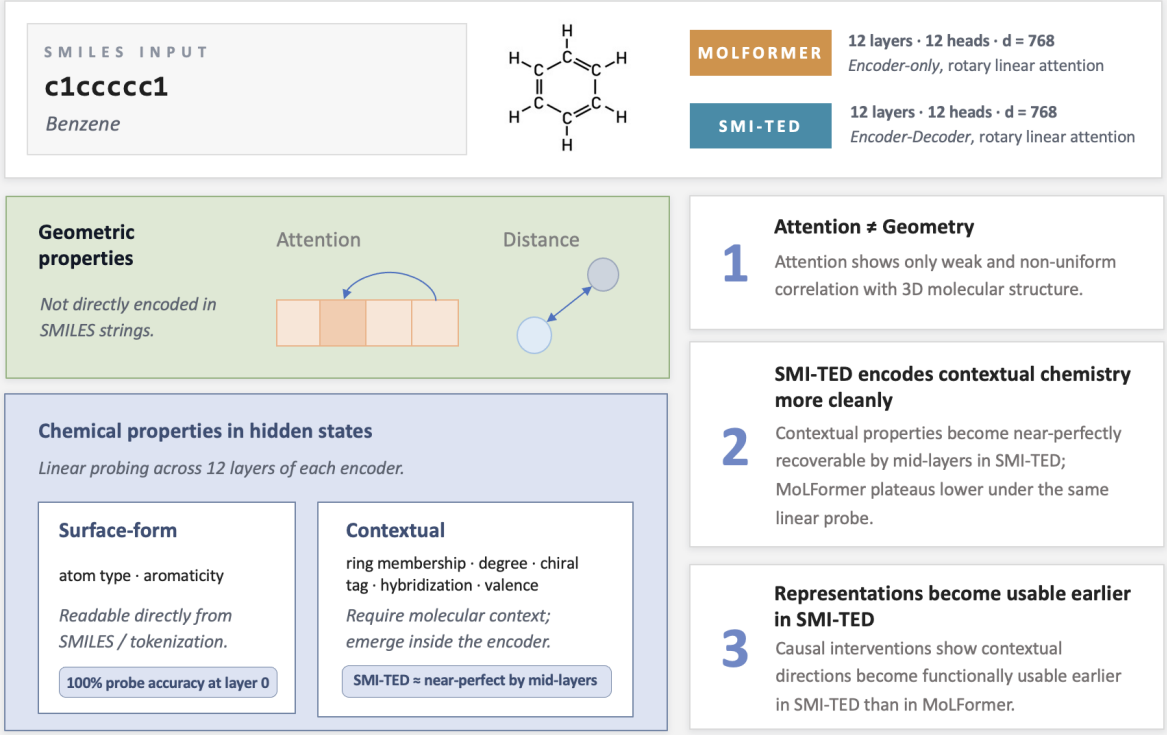

Figure 1. Overview of our mechanistic analysis of chemical foundation models. We compare MoLFormer and SMI-TED using three complementary experiments: (1) layerwise attention–distance correlation, (2) linear probing of eight per-atom RDKit-derived chemical properties, and (3) causal intervention along probe directions.

2025) to distinguish surface-level token information from contextual chemical structure; and (3) causal interventions along probe directions to test whether decoded properties are actively used during computation beyond being linearly recoverable (Figure 1).

We find that the two models, despite their architectural similarity, organize chemical information in different ways. Attention in both encoders only weakly tracks 3D molecular geometry, with highly non-uniform Spearman correlations below 0.3 across all layer-head pairs, which contrast with prior claims that MoLFormer attention encodes spatial structure (Ross et al., 2022). Surface-level properties directly encoded in SMILES tokens, such as atom type and aromaticity, reach nearly 100% probe accuracy at the embedding layer in both models. In contrast, contextual properties that require aggregating information across the molecular graph emerge later in the encoder, with SMI-TED encoding them more cleanly than MoLFormer. Causal interventions further reveal that linear recoverability and functional use are not equivalent: contextual properties become causally usable by mid-depth in SMI-TED but only near the top of the encoder in MoLFormer.[1]

To summarize, our main contributions are as follows:

- We perform a mechanistic interpretability study of MoLFormer and SMI-TED, analyzing how they represent chemical structure and properties from SMILES strings. Revisiting prior claims that molecular attention captures spatial structure, we show that attention patterns correlate only weakly with true 3D interatomic distance despite strong downstream chemistry performance.
- Using layerwise probing and causal interventions across RDKit-derived atomic properties, we show that surface-level properties are directly recoverable from token embeddings, while contextual chemical properties are encoded much more cleanly and become causally usable earlier in SMI-TED than in MoLFormer.
- Our results suggest that linear decodability and causal usage capture different aspects of representation learning in molecular foundation models, motivating more careful interpretation of probing results in scientific domains.

2. Background

MoLFormer and SMI-TED. We study two SMILES-based chemistry foundation models: MoLFormer (Ross et al., 2022) and SMI-TED (Soares et al., 2024). Both are transformer models that use rotary linear attention and were pretrained without explicit geometric supervision, making them a useful comparison pair for studying how different

[1]Code Available: https://github.com/Elena-Baixy/smimolm

pretraining objectives and architectural choices shape internal chemical representations.

MOLFORMER is an *encoder-only* model trained with a masked language modeling objective. It produces token-level hidden states and a pooled molecular representation for downstream prediction. In the original paper, MOLFORMER is evaluated on molecular property prediction, molecular retrieval, and quantum-chemical benchmarks including MoleculeNet and ADMET-related tasks, where it achieves strong downstream performance. The paper additionally analyzes attention patterns and embedding structure to argue that the learned representations capture aspects of molecular 3D spatial structure and chemical similarity (Ross et al., 2022; Wu et al., 2018).

SMI-TED uses an *encoder–decoder* design with both masked token prediction and sequence reconstruction. Its latent representation is therefore trained not only to support token-level inference, but also to retain enough information to reconstruct the input SMILES. The original SMI-TED is evaluated on similar tasks as MOLFORMER (Soares et al., 2024; Wu et al., 2018).

Chemical Properties and SMILES Strings. Given a molecule, each atom is annotated with eight properties derived from RDKit (Landrum et al., 2025), an open-source cheminformatics toolkit. These properties should be understood as local chemical descriptors that capture chemically meaningful structural information at the atom level, rather than full molecule-level behavior. We distinguish these descriptors from geometric properties, such as interatomic distance, which are not explicitly encoded in SMILES strings.

Within the atom-level descriptors, we distinguish between **surface-level properties**, whose labels are directly readable from individual SMILES tokens, and **contextual properties**, which require the encoder to aggregate information across the molecular graph. Atom type and aromaticity are surface-level since element symbols are encoded as distinct tokens, and aromatic atoms are marked by lowercase letters in SMILES notation. By contrast, hybridization, degree, total valence, ring membership, formal charge, and chiral tag cannot be determined from individual SMILES tokens in isolation. These contextual properties therefore provide a stronger test of whether the encoder has learned chemically meaningful structure beyond surface-level tokenization patterns. Detailed definitions of all properties are provided in Appendix A.1.

3. Methods

Both MOLFORMER (Ross et al., 2022) and SMI-TED (Soares et al., 2024) use 12-layer, 12-head transformer encoders with hidden dimension $d = 768$. We run the same set of analyses on both models to compare whether their learned representations encode molecular structure in similar ways. The attention-distance experiment tests whether attention heads directly track 3D spatial proximity, while the probing experiment tests whether atom-level chemical properties are linearly recoverable from hidden states. Throughout, we use RDKit (Landrum et al., 2025) to access molecular properties and generate 3D conformers.

3.1. Attention-Distance Correlation.

For each layer–head pair, we compare attention weights between **heavy** (non-Hydrogen) atom pairs $A_{ij}^{(l,h)}$ to their corresponding 3D Euclidean distances D_{ij} (measured in Å).[2] We compute the Spearman correlation ρ between attention weights and inverse distance separately for every molecule, then average the resulting correlations across the sample. More detailed setup is in Appendix B.

We additionally stratify atom pairs by distance into *short* ($D_{ij} \leq 2$ Å), *medium* ($2 < D_{ij} \leq 4$ Å), and *long* ($D_{ij} > 4$ Å), and repeat the analysis within each group.

3.2. Probing Chemical Properties

Probing. Let $\mathbf{h}_l^{(i)} \in \mathbb{R}^d$ denote the residual-stream hidden state at the i-th atom-token position at layer l, where d=768 and $l \in \{0, 1, \ldots, 12\}$ (layer 0 is the embedding output). We fit a linear probe $\mathbf{h} \mapsto \mathbf{W}\mathbf{h} + \mathbf{b}$ with $\mathbf{W} \in \mathbb{R}^{C_\tau \times d}$ and $\mathbf{b} \in \mathbb{R}^{C_\tau}$, where C_τ is the number of classes for task τ. We compare the linear probe to two references: (i) a *random-input* linear probe on a Gaussian feature matrix $X_{ij} \sim \mathcal{N}(0, 1)$ in place of the real hidden states (controls for probe capacity) and (ii) a *majority-vote* baseline that always predicts the most frequent training-set class. We report test accuracy and macro-F1.

Causal Intervention. For a given task, choice of source class c, target class c', and steer layer ℓ, we shift every atom-token position whose baseline prediction is c by $\alpha \cdot (\mathbf{W}_{c'} - \mathbf{W}_c)/\boldsymbol{\sigma}$ at layer ℓ, then run the rest of the encoder and collect the layer-12 prediction. *Intervention success rate* is the fraction of those positions whose post-intervention layer-12 prediction equals c'. We run a separate single-layer intervention at each candidate $\ell \in \{1, \ldots, 12\}$, using the probe trained at that layer's hidden states to define both the source-class detector and the steering direction, and report the success rate at the best-performing α on held-out molecules. More detailed setup is in Appendix C.

Tasks. We probe eight per-atom RDKit-derived properties on QM9 (Ramakrishnan et al., 2014): atom type (C_τ=4), hy-

[2]Since SMILES strings do not contain explicit 3D geometric information, we use ETKDGv3 and MMFF94 (modules from RDKit) to generate a 3D coordinate $\mathbf{r}_i = (x_i, y_i, z_i)$ for each atom. This procedure is used only as a preprocessing step; no additional model training is performed.

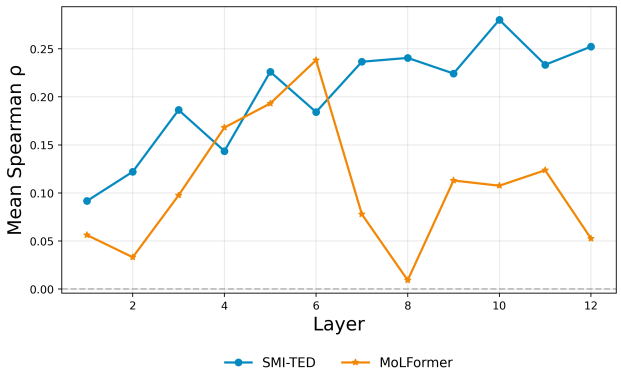

Figure 2. Mean attention-distance correlation per layer for MoLFormer and SMI-TED. Both models exhibit low correlations with SMI-TED performing marginally better across layers.

bridization (C_τ=3), chiral tag (C_τ=3), aromaticity (C_τ=2), ring membership (C_τ=2), degree (C_τ=4), formal charge (C_τ=3), and total valence (C_τ=4). Integer-valued properties (degree, valence) are treated as discrete classes rather than ordinal regression. Atom-token alignment, train/test split, optimization hyperparameters, and standardization are detailed in Appendix C.

4. Results

We study how molecular language models represent chemical structure at both the attention and hidden-state level. Revisiting prior claims that MoLFormer attention captures spatial relationships between atoms, we find that attention correlates only weakly with true interatomic distance. Probing further reveals that SMI-TED encodes contextual chemical properties more strongly than MoLFormer, and that these representations are actively used during computation rather than merely correlated with the target properties.

4.1. Attention-Distance Correlation

Attention weakly reflects molecular geometry. Although both models exhibit consistently positive mean correlations, all layer-wise mean Spearman correlations remain below $\rho \approx 0.3$ (Figure 2), suggesting that 3D geometry is only weakly encoded in attention patterns for both models.

Overall, SMI-TED exhibits consistently stronger correlations than MoLFormer. Notably, the gradual increase in correlation across layers in SMI-TED suggests that geometric information emerges progressively through the transformer hierarchy rather than being encoded uniformly at early layers. In contrast, MoLFormer peaks at layer 6 before declining in later layers, suggesting that later MoLFormer layers may instead capture more abstract or non-spatial molecular relationships.

Attention exhibits local but non-uniform geometry alignment. Across both models, attention–distance correlations are strongest for short- and medium-range atomic interactions and weaken at longer distances. SMI-TED exhibits consistently stronger geometric alignment than MoLFormer (Figure 3). In the short-distance group, many SMI-TED heads attain correlations between 0.40 and 0.50, whereas the maximum short-range correlation in MoLFormer is 0.25. Similarly, medium-distance correlations in SMI-TED frequently fall between 0.35 and 0.45, while MoLFormer medium-distance correlations peak at 0.29. These results suggest that SMI-TED preserves spatial structure more consistently than MoLFormer.

Furthermore, several heads in both models exhibit negative correlations, suggesting that some heads capture chemical features not reflected by Euclidean distance alone, which may indicate that the models do not associate distance in the way human chemists do.

Comparison with prior MoLFormer findings. Our findings contrast with prior claims in MoLFormer that attention strongly aligns with interatomic proximity. While both analyses compare attention weights with inverse pairwise distances, the original work aggregates attention across heads before computing cosine similarity with the inverse-distance matrix, potentially obscuring substantial head-level heterogeneity. In addition, cosine similarity measures global directional alignment, whereas our per-head Spearman analysis directly measures monotonic association between attention strength and spatial proximity. Our results instead suggest that geometric alignment is weak, highly non-uniform across heads, and largely absent for long-range interactions (Ross et al., 2022).

4.2. Linear probing of chemical properties

Surface-level properties are recoverable at the embedding in both models. The probe accuracy for atom type (Figure 4e) and aromaticity (Figure 4a), two surface-level properties, is 1.0 at the embedding layer (ℓ=0) in both MoLFormer and SMI-TED and stays flat across all 13 layers, indicating these properties are directly encoded in the embedding and unchanged by any encoder computation.

SMI-TED encodes contextual properties more cleanly than MoLFormer. The clearest cross-model gap appears on the contextual properties: ring membership, degree, chiral tag, hybridization, and total valence (Figure 4b, 4c, 4d, 4f, 4h). The same C token can belong to a ring or not, can have different degree, and can carry different stereochemistry depending on its surroundings, so a probe that achieves high accuracy on these tasks must read information that the encoder put there from context. SMI-TED reaches near-perfect accuracy (≥ 0.95) on all five contextual properties. Total valence and chiral tag are already above this thresh-

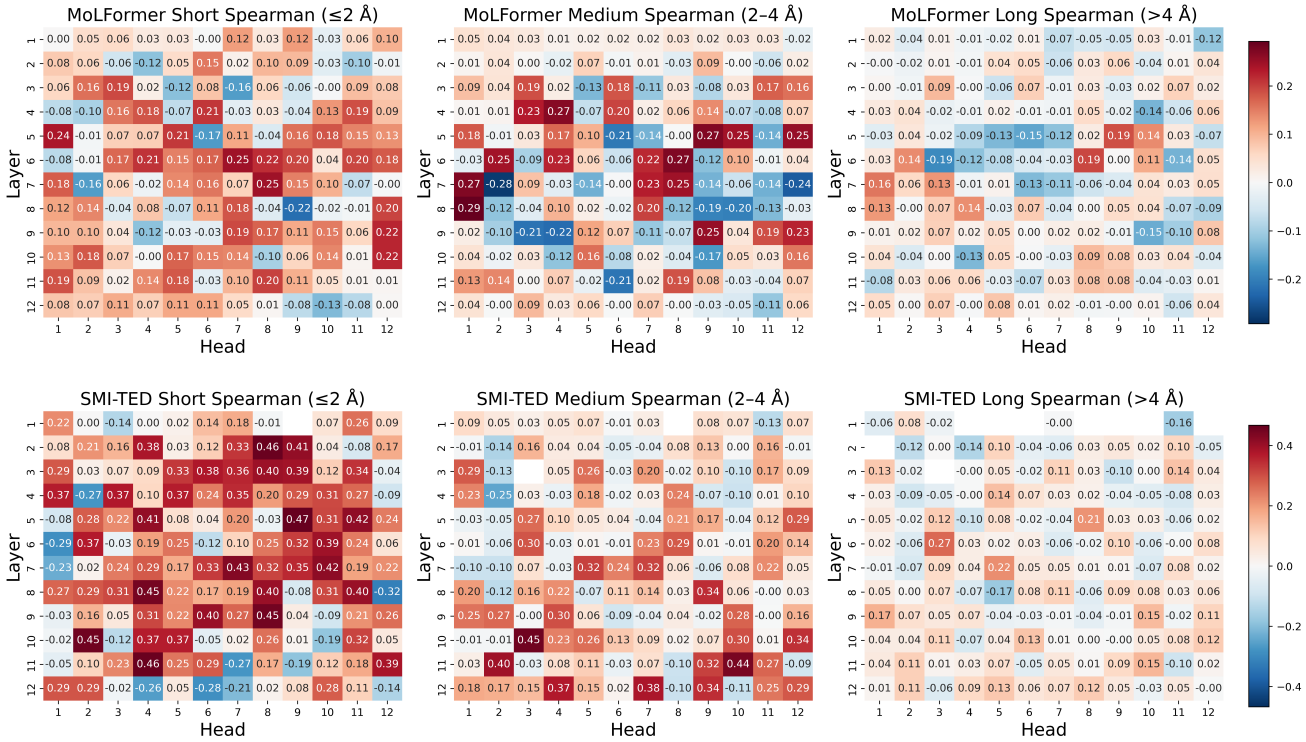

Figure 3. Attention-distance correlation matrices per layer, head pair for MoLFormer and SMI-TED stratified by short, medium, long distance (measured in Å). Both models exhibit stronger correlations at short and medium distances, with correlations weakening considerably for long-range atomic interactions.

old at the embedding. Other properties also converge in the early layers. MoLFormer's curves climb less steeply and plateau several points below. Ring membership levels off around 0.82, hybridization around 0.84, total valence near 0.97, and degree at ~ 0.58 (only 0.15 above the 0.43 majority baseline). MoLFormer's probe accuracy stays at ~ 0.79 across all 13 layers for chiral tag, essentially the 0.795 majority-class baseline. This suggests differences in how the two models encode contextual chemical properties.

4.3. Causal intervention on chemical properties

Aromaticity (Figure 5a) and formal charge (Figure 5g) show intervention success ≈ 1.0 at every depth in both models, indicating that these properties are usable from the embedding onward and remain available throughout the encoder. This matches the probing results: when a property is already explicit in the representation, both models can route through it at essentially every layer.

The contextual properties show a different pattern. In SMI-TED, intervention success rises rapidly and typically saturates by mid-depth, around layers 6–7. In MoLFormer, the same properties improve more gradually and often do not reach ceiling until the top few layers, around layers 9–12. The two models therefore appear to use the same chemical concepts differently across depth. In SMI-TED, contextual information becomes a usable computational feature relatively early, whereas in MoLFormer it is incorporated more gradually and often only becomes fully usable near the top of the encoder. This suggests that SMI-TED organizes contextual chemistry into actionable directions earlier, which may be advantageous for more complex downstream tasks that require later layers to build on ring structure, valence, or stereochemistry as intermediate features rather than only read them out at the end.

In SMI-TED, the success curves additionally show a sharp dip at layer 11 across multiple properties. Sweeping α across both a fine grid in $[0, 1]$ and a wide grid up to $\alpha = 300$ at layer 11 never recovers the L10 / L12 ceiling for ring membership, degree, chiral tag, hybridization, or formal charge (Table 1). We therefore treat this effect cautiously: it is consistent with prior concerns that late-layer readouts and interventions can become brittle or misaligned with the model's own computation (Park et al., 2024; Belrose et al., 2023; Heimersheim & Nanda, 2024), rather than taking it as a robust model-specific scientific effect.

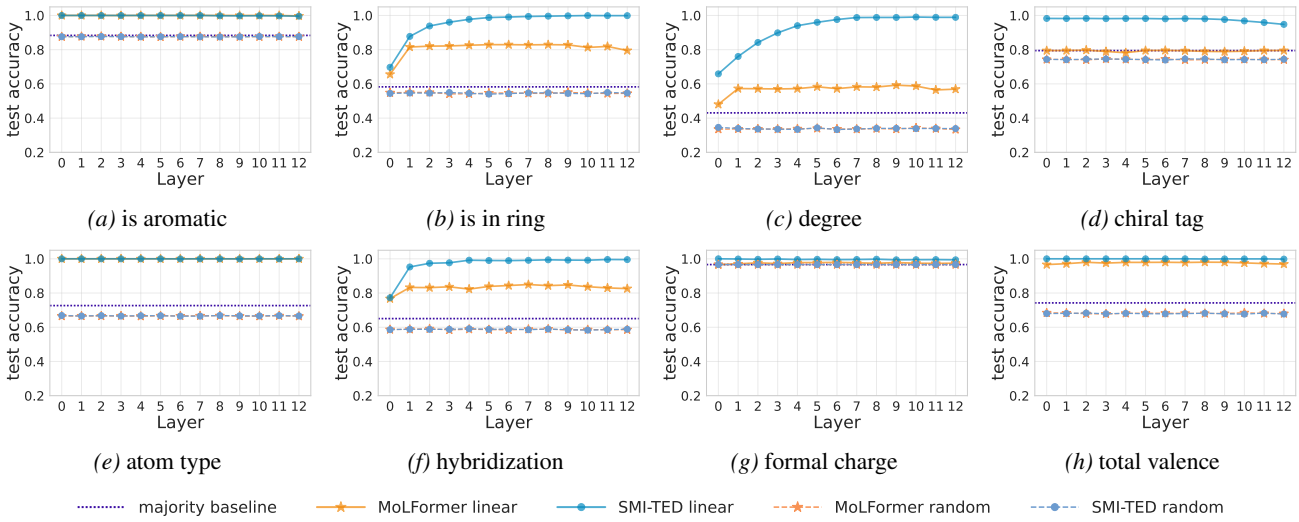

Figure 4. Linear-probe test accuracy per layer on eight per-atom QM9 properties. SMI-TED reaches near-perfect accuracy by mid-layers on the contextual properties (ring membership, degree, chiral tag, hybridization, total valence), while MOLFORMER plateaus several points below; the surface-form properties (atom type, aromaticity) saturate at the embedding in both models.

5. Related Work

Downstream applications of MOLFORMER and SMI-TED. Both encoders we study are widely used as fixed feature extractors or finetuning backbones for downstream chemistry tasks. MOLFORMER has been applied to molecular property prediction across the MoleculeNet benchmarks (Ross et al., 2022; Wu et al., 2018) and adapted to ADMET and bioactivity prediction with parameter-efficient finetuning (Harod et al., 2025). SMI-TED has been used as a backbone for multi-scale battery materials and electrolyte discovery (Sharma et al., 2025), and SMILES encoders of this family appear as components in multimodal property predictors that combine text and graph features (Rollins et al., 2024) and in scaling studies of chemical foundation models (Cai et al., 2026). These applications treat the encoders as black-box feature sources, which motivates our question of where in their layer stacks chemical information is actually represented and used.

Interpreting scientific foundation models. On chemistry, Pinto (2025a) find that intermediate layers of SMILES encoders outperform the final layer for downstream tasks, Cohen et al. (2025) apply sparse autoencoders to SMI-TED and steer generation by intervening on recovered features, and Varadi et al. (2025); Müller et al. (2025); Hu et al. (2024) interpret attention heads, prediction circuits, and fragment-level features of other SMILES models. A parallel line of work targets biological foundation models: Lu et al. (2026) use counterfactual interventions on ESMFold latents to trace how the model builds spatial structure from sequence, Goodfire AI (2026b) probe Evo 2 to generate interpretable predictions for millions of genetic variants, and Goodfire AI (2026a) use sparse autoencoders on an epigenetic foundation model to identify a novel Alzheimer's biomarker class. Building on this line of work, we pair layerwise linear probing with causal intervention on probe directions in MOLFORMER and SMI-TED, which lets us separate information that is merely linearly recoverable from information the encoder actually uses during computation.

6. Discussion

Our results show that strong downstream performance in molecular foundation models does not necessarily correspond to spatially grounded or uniformly organized chemical representations. Although both MOLFORMER and SMI-TED encode chemically relevant information, they differ in how cleanly that information is represented and when it becomes causally usable during computation. More broadly, our findings suggest that interpretability methods developed for language models may not transfer directly to scientific foundation models, motivating tools that better account for scientific data, architectures, and reasoning objectives.

6.1. Chemical information without human-like organization

Strong performance on chemistry tasks is often interpreted as evidence that a model has learned meaningful chemical structure. However, our results show that high downstream accuracy does not necessarily imply structured or human-like chemical representations. Although prior work suggested that attention in molecular language models captures spatial relationships between atoms, we find that attention exhibits only weak alignment with true 3D geometry. Instead, chemically relevant information is primarily rep-

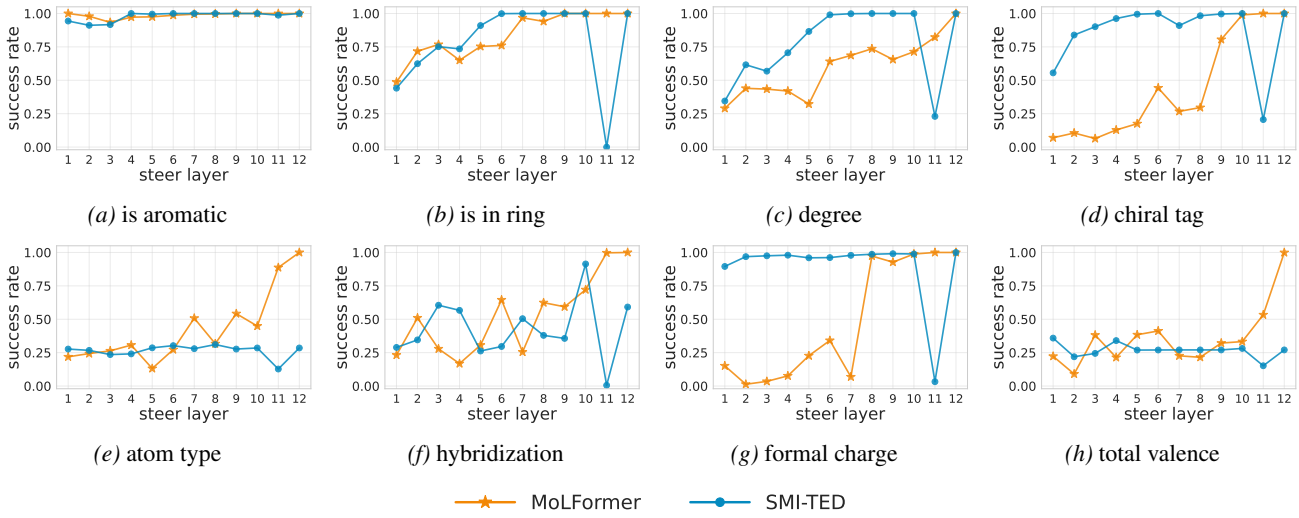

Figure 5. Intervention success rate per steer layer on eight per-atom QM9 properties. On the contextual properties, SMI-TED's success rises rapidly and saturates at 1.0 by mid-layers, while MoLFormer's climbs gradually and only reaches 1.0 near the top of the encoder, indicating different depths of usage.

resented in the hidden states, and its organization differs across models. We further show that some properties become linearly recoverable before they are causally used by later layers, indicating that representation accessibility and computational use emerge at different stages of the network. In particular, MoLFormer and SMI-TED achieve similarly strong downstream performance on overlapping MoleculeNet and ADMET-style benchmarks (Ross et al., 2022; Soares et al., 2024; Wu et al., 2018; Harod et al., 2025; Sharma et al., 2025). However, they differ in how contextual properties are encoded and how early those encodings become causally usable. These results suggest that similar predictive performance can arise from different internal representations.

These differences matter because downstream scientific applications often require more than accurate prediction alone. In practice, we may want molecular models to organize chemical knowledge in ways that support compositional reasoning, reliable transfer, and faithful interpretation. Tasks such as ADMET prediction, molecular binding, and reactivity frequently require combining multiple local and non-local chemical constraints across several computational steps. Representations that become accessible and causally usable earlier in the network can serve as intermediate features for later computation. This may be especially important in transfer settings where molecular encoders are used through frozen embeddings, shallow predictors, parameter-efficient adaptation, or intermediate-layer readouts rather than fully end-to-end retraining, so that the downstream model can still capture the information (Honda et al., 2019; Rollins et al., 2024; Cai et al., 2026; Harod et al., 2025; Pinto, 2025b). Under these constraints, downstream performance may depend not only on whether chemical information is present somewhere in the model, but also on how cleanly and causally that information is organized.

6.2. Limitations of probing in molecular models

Related to the previous section, our results suggest that scientific foundation models may learn and organize domain knowledge in ways that differ from human intuition, which also changes how interpretability tools should be applied to these models.

Standard probing methods can become difficult to interpret in this setting due to biased scientific datasets, domain-specific tokenization, different training objectives, and the distinct scientific questions we care about. Chemistry atom-level datasets are naturally biased toward certain classes because most atoms in nature are carbon, most carbons are non-aromatic, and most chiral centers carry a single configuration. This might skew the representation space so that a probe can find a direction the encoder does not actually use. Our results expose this gap in three concrete forms. First, on atom type, both MoLFormer and SMI-TED achieve near-perfect probe accuracy at every layer, yet intervention along the probe direction succeeds in MoLFormer only at the very top of the encoder and never in SMI-TED for atom type (Figure 5e). This aligns with what Park et al. (2024) suggested: linear directions in a model's *output* space need not coincide with the *input* directions later layers use. Second, MoLFormer's chiral-tag probe stays at the 0.795 majority baseline at every depth, yet steering still flips the readout at the top of the encoder. With no real linear feature to perturb, this nominal success is the false-positive failure mode that Hewitt & Liang (2019) introduced con-

trol tasks to detect, and an instance of the encoded-versus-used bug catalogued in NLP probing (Antverg & Belinkov, 2022; Elazar et al., 2021). Third, the L11 dip we report in SMI-TED shows that even a clean probe direction can live in a basis the next encoder block does not read. This layer-specific brittleness is consistent with prior reports on late-layer readouts and LayerNorm-mediated self-repair in transformers (Belrose et al., 2023; Heimersheim & Nanda, 2024).

This aligns with broader literature that probing accuracy is a noisy proxy for what the model actually uses (Belinkov, 2022; Pimentel et al., 2020), and probes decoding a feature need not identify the directions the model relies on (Lasri et al., 2022), which can be amplified by the nature of scientific datasets, tokenization and model design.

6.3. Challenges for interpretability tools for scientific foundation models

Our results suggest that it is possible that interpretability methods developed for language models do not transfer directly to molecular models, even in general scientific foundation models. Scientific models often operate on different input structures, training objectives, and reasoning tasks, which changes both what representations emerge and how interpretability results should be interpreted.

One challenge is *tokenizer leakage*. In molecular models such as MoLFormer and SMI-TED, SMILES tokens directly encode atom identity, aromaticity, and parts of stereochemistry. Similar issues appear in biological foundation models where amino-acid or genomic tokenizations partially expose biochemical structure (Rives et al., 2021; Lin et al., 2023b). As a result, high probe accuracy may reflect explicit information in the input representation rather than learned compositional structure. Scientific datasets also often exhibit strong class imbalance and structural bias. Rare chemical, biological, or materials properties may occupy only a small fraction of the data, making standard probing metrics such as accuracy difficult to interpret. High probe performance can therefore reflect majority-class behavior rather than meaningful representation learning.

Another challenge is architectural and objective mismatch. Unlike standard autoregressive language models, scientific foundation models frequently incorporate specialized inductive biases, reconstruction objectives, geometric supervision, or multimodal structure (Ross et al., 2022; Soares et al., 2024; Jumper et al., 2021). These differences can alter how information is represented and propagated across layers.

Finally, scientific applications often require more than accurate prediction alone. We may additionally care whether models organize knowledge in ways that support compositional reasoning, physical consistency, extrapolation, and faithful scientific interpretation. A feature being linearly decodable may therefore be insufficient; understanding whether it is causally used and how it interacts with other scientific concepts may be equally important.

Together, these challenges suggest that scientific foundation models require interpretability methods tailored to scientific representations and reasoning, rather than direct transfer of tools developed for natural language models.

7. Conclusion

Taken together, the attention and probing results suggest that chemical information is not represented uniformly across mechanisms or layers. Attention exhibits weak and highly selective alignment with molecular geometry, while probing and causal intervention show that chemically meaningful properties are strongly represented within the residual stream but become usable at different depths across models. More broadly, our results move beyond asking whether chemical information is present toward understanding how it is internally organized and used during computation.

Future directions. Several directions remain open for understanding chemical representations in molecular foundation models. While probing identifies which properties are recoverable, it does not explain how chemical features are organized across layers or whether they occupy shared subspaces, motivating future work on representation geometry, feature overlap, and superposition using methods such as sparse autoencoders or representation similarity analysis. Our interventions further show when a representation becomes causally usable, but not how later components consume or combine this information into more global molecular reasoning, suggesting future analysis of attention circuits and residual-stream interactions. More broadly, although MoLFormer and SMI-TED achieve similar downstream performance, they differ in how and when chemical information becomes usable, raising the broader question of how pretraining objectives and architectural choices shape internal scientific representations.

Acknowledgements

This project is inspired by initial experiments conducted by Neurico (Haokun Liu, 2025). This work is in part supported by the University of Chicago Novel Intelligence Research Initiative and AI research pillars.

References

Antverg, O. and Belinkov, Y. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*, 2022.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112.

Cai, F., Zacour, K., Zhu, T., Tzeng, T.-R., Duan, Y., Liu, L., Pilla, S., Li, G., and Luo, F. Chemfm as a scaling law guided foundation model pre-trained on informative chemicals. *Communications Chemistry*, 9(3), 2026. doi: 10.1038/s42004-025-01793-8. URL https://www.nature.com/articles/s42004-025-01793-8.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., and Weller, A. Rethinking attention with Performers. In *International Conference on Learning Representations*, 2021.

Cohen, R., Hasson, D., and Tanovic, M. Unveiling latent knowledge in chemistry language models through sparse autoencoders. *arXiv preprint arXiv:2512.08077*, 2025.

Elazar, Y., Ravfogel, S., Jacovi, A., and Goldberg, Y. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Goodfire AI. Using interpretability to identify a novel class of Alzheimer's biomarkers, 2026a. URL https://www.goodfire.ai/research/interpretability-for-alzheimers-detection. Goodfire AI Research.

Goodfire AI. Explaining 4.2 million genetic variants with state-of-the-art, interpretable predictions, 2026b. URL https://www.goodfire.ai/research/evee-explaining-genetic-variants. Goodfire AI Research.

Haokun Liu, C. T. Neurico: Autonomous research framework, 2025. URL https://github.com/ChicagoHAI/neurico.

Harod, H., Agrawal, K., Simeone, I., Ceccarelli, M., Gupta, S., and Mall, R. Effichem: Efficient adaptation of chemical language models for molecular property prediction, 2025. URL https://chemrxiv.org/engage/chemrxiv/article-details/68b7f02e728bf9025ef4fcfc. ChemRxiv preprint.

Heimersheim, S. and Nanda, N. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, 2019. Association for Computational Linguistics.

Honda, S., Shi, S., and Ueda, H. R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery, 2019. URL https://arxiv.org/abs/1911.04738.

Hu, X., Liu, G., Zhao, Y., and Zhang, H. Empirical evidence for the fragment level understanding on drug molecular structure of LLMs. *arXiv preprint arXiv:2401.07657*, 2024.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L. Y., Zhang, J., and Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Landrum, G. et al. Rdkit: Open-source cheminformatics. https://www.rdkit.org/, 2025. Accessed: 2026-05-08.

Lasri, K., Lenci, A., and Poibeau, T. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8818–8831, Dublin, Ireland, 2022. Association for Computational Linguistics.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a. doi: 10.1126/science.ade2574.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023b.

Lu, K., Brinkmann, J., Huber, S., Mueller, A., Belinkov, Y., Bau, D., and Wendler, C. Mechanisms of AI protein folding in ESMFold, 2026. URL https://arxiv.org/abs/2602.06020.

Müller, A., Cardenas Cartagena, D., and Pollice, R. Uncovering internal prediction mechanisms of transformer-based chemical foundation models. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-d1j8k.

Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D. B., Bartie, L. J., Thomas, A. W., King, S. H., Brixi, G., Sullivan, J., Ng, M. Y., Lewis, A., Lou, A., Ermon, S., Baccus, S. A., Hernandez-Boussard, T., Ré, C., Hsu, P. D., and Hie, B. L. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723): eado9336, 2024. doi: 10.1126/science.ado9336.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Pimentel, T., Valvoda, J., Hall Maudslay, R., Zmigrod, R., Williams, A., and Cotterell, R. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622. Association for Computational Linguistics, 2020.

Pinto, L. Superior molecular representations from intermediate encoder layers. *arXiv preprint arXiv:2506.06443*, 2025a.

Pinto, L. Unlocking chemical insights: Superior molecular representations from intermediate encoder layers, 2025b. URL https://arxiv.org/abs/2506.06443.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014. doi: 10.1038/sdata.2014.22.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118 (15):e2016239118, 2021.

Rollins, Z. A., Cheng, A. C., and Metwally, E. Molprop: Molecular property prediction with multimodal language and graph fusion. *Journal of Cheminformatics*, 16 (56), 2024. doi: 10.1186/s13321-024-00846-9. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-024-00846-9.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties, 2022. URL https://arxiv.org/abs/2106.09553.

Sharma, V., Giammona, M. J., Tek, A., Zohair, M., Park, N. H., Erdmann, T., Soares, E., Sundberg, L., Nguyen, K., Na, Y.-H., et al. Foundation models enabling multi-scale battery materials discovery: From molecules to devices. In *AI for Accelerated Materials Design-NeurIPS 2025*, 2025.

Soares, E., Brazil, E. V., Shirasuna, V. Y., Zubarev, D., Cerqueira, R., and Schmidt, K. Smi-ted: A large-scale foundation model for materials and chemistry. 2024.

Sterling, T. and Irwin, J. J. ZINC 15—ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

Varadi, K., Marosi, M., and Antal, P. Circuits, features, and heuristics in molecular transformers. *arXiv preprint arXiv:2512.09757*, 2025.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

A. Models and Atomic Properties

MoLFormer. MoLFormer (Ross et al., 2022) is an encoder-only transformer that applies rotary linear attention to SMILES token sequences and produces both token-level hidden states and a pooled molecular representation. The public release we study has a 12-layer encoder with hidden size 768 and 12 attention heads, and a vocabulary of 2,362 tokens. The encoder is pretrained on roughly 1.1 billion molecules from PubChem (Kim et al., 2023) and ZINC (Sterling & Irwin, 2015) under a masked language modeling objective, with no explicit supervision on geometry or molecular properties. Each token is embedded in a 768-dimensional space, and sequences are capped at 202 tokens, which accommodates over 99.4% of molecules in the training corpus. MoLFormer replaces standard softmax attention with a positive-feature-map approximation (Choromanski et al., 2021), combined with rotary position encodings R_m applied to queries and keys at each position m. The resulting attention at position m is

$$\text{Attention}_m(Q, K, V) = \frac{\sum_{n=1}^{N} \langle \varphi(R_m q_m),\, \varphi(R_n k_n) \rangle\, v_n}{\sum_{n=1}^{N} \langle \varphi(R_m q_m),\, \varphi(R_n k_n) \rangle}, \tag{1}$$

where φ is a random feature map that enables linear-time attention and avoids the quadratic cost of standard softmax attention.

SMI-TED. SMI-TED (Soares et al., 2024) uses the same general encoder backbone (12 layers, 768-dimensional hidden states, 12 attention heads, rotary linear attention) but adds an encoder–decoder bottleneck trained jointly with token-level prediction and sequence reconstruction. The encoder maps token embeddings $\mathbf{x}$ into a compressed latent representation $\mathbf{z}$ via

$$\mathbf{z} = \text{LayerNorm}(\text{GELU}(\mathbf{x}W_1 + b_1))\, W_2, \tag{2}$$

and reconstructs the token sequence from $\mathbf{z}$ through a symmetric immersion followed by a language decoder layer. The encoder contains 47M parameters and the decoder 242M parameters, totalling 289M parameters. This bottleneck imposes an additional constraint on the latent space: it must retain enough information to support both downstream prediction and recovery of the input SMILES, which makes SMI-TED a useful comparison point for studying whether different pretraining designs lead to different internal organization of chemical information.

A.1. Atomic Properties Derived from RDKit

Each atom is annotated with eight RDKit-derived properties, falling into three categories of how readable they are from the SMILES string itself.

Surface-form (directly recoverable from SMILES tokenization). *Atom type* is the chemical element identity of the atom. Element symbols are explicitly encoded in the SMILES string and assigned distinct token ids by the tokenizer, so atom type is directly readable. *Aromaticity* indicates whether an atom belongs to an aromatic system; aromatic atoms are written with lowercase letters in SMILES notation, so aromaticity is also surface-form.

Partially signaled by SMILES. *Formal charge* is the net charge assigned to an atom relative to its neutral state. It is sometimes indicated by bracket notation, but consistent labels require RDKit standardization across representations. *Chiral tag* encodes the stereochemical configuration at a chiral center. SMILES partially signals chirality with `@` and `@@`, but canonical assignment requires stereochemical perception by RDKit.

Contextual (require cheminformatics computation). *Hybridization* describes the orbital hybridization state of an atom; aromatic atoms implicitly carry sp^2 character via lowercase notation, but hybridization for non-aromatic atoms requires full valence resolution. *Degree* is the number of directly bonded heavy-atom neighbors of each atom; while bond connectivity is encoded in SMILES, degree must be computed by parsing the molecular graph since implicit hydrogens complicate direct reading. *Total valence* counts all bonds formed by an atom, including bonds to implicit hydrogens, which SMILES typically suppresses. *Ring membership* indicates whether an atom belongs to one or more rings; SMILES encodes ring closures through numbered bond descriptors, but assigning ring membership to each atom (especially in fused or bridged polycyclic systems) requires explicit ring perception.

The five contextual properties (and partially-signaled formal charge and chiral tag) cannot be determined from raw SMILES alone, and therefore serve as the most informative targets for evaluating how much structural information a model has learned beyond its tokenization. We additionally consider one geometric property, **interatomic distance**, in our attention analysis (Appendix B); this property is not encoded in the SMILES string at all and must be estimated from a generated 3D conformer.

B. Attention–Distance Details

We test whether attention preferentially connects atoms that are spatially close in 3D. For each layer–head pair (l, h), we compute the Spearman correlation between attention weights and inverse interatomic distance across all unique heavy-atom pairs $i < j$, then average these correlations across molecules. We randomly sample 1000 SMILES strings from the QM9 training split, excluding molecules with fewer than three heavy atoms, and apply the same procedure to both models.

Attention extraction. Both MoLFormer and SMI-TED use the FAVOR+ linear-attention approximation (Choromanski et al., 2021), which does not explicitly materialize the full attention matrix. We therefore capture the query and key tensors $\mathbf{q}_i^{(l,h)}$ and $\mathbf{k}_j^{(l,h)}$ at every layer and head, reconstruct the effective attention weights post hoc as $A_{ij}^{(l,h)} \propto \langle \varphi(\mathbf{q}_i^{(l,h)}), \varphi(\mathbf{k}_j^{(l,h)}) \rangle$, and apply row normalization. We restrict attention to atom-token rows and columns, discarding non-atomic tokens.

Conformer generation. Because SMILES strings encode molecular connectivity but not 3D geometry, we generate a single low-energy conformer for each molecule using RDKit ETKDGv3 followed by MMFF94 optimization, assigning each heavy atom coordinates $\mathbf{r}_i = (x_i, y_i, z_i)$.

Distance correlation. From the optimized conformer, we compute the pairwise Euclidean distance matrix $D_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|_2$, form the inverse-distance matrix D^{-1}, and compute the Spearman correlation between $\{A_{ij}^{(l,h)}\}$ and $\{1/D_{ij}\}$ over all unique heavy-atom pairs $i < j$. We additionally stratify correlations by short-range (≤ 2 Å), medium-range (2–4 Å), and long-range (> 4 Å) interactions to localize where geometric signal is concentrated.

Summary. Both models exhibit only weak overall alignment between attention and 3D geometry, with the strongest correlations concentrated in short- and medium-range atom pairs.

C. Probing and Causal Intervention Details

C.1. Probe training and data

Optimization. Each linear probe is a single fully-connected layer $\mathbf{W} \in \mathbb{R}^{C_\tau \times d}$, $\mathbf{b} \in \mathbb{R}^{C_\tau}$ trained from random initialization for every (task, layer) pair under cross-entropy loss. We use AdamW with learning rate 10^{-3}, weight decay 10^{-2}, batch size 256, 50 epochs, and gradient norm clipping at 5.0. The MLP probe and the random-features probe (described below) use the same optimizer, schedule, and standardization as the linear probe.

Data and split. We sample 1000 SMILES uniformly at random from the QM9 training split with seed 42. After RDKit parsing and atom-token alignment, approximately 980 molecules and ~ 8.6k atom-token positions remain. We use a molecule-level $80/20$ split so that no atom from a held-out molecule appears in training, yielding ~ 6.9k training and ~ 1.7k test atom-token positions. Class sets are fit on training labels only; held-out atoms whose label does not appear in training count as misclassifications and are excluded from macro-F1.

Per-atom hidden-state extraction. For each SMILES we run a forward pass with `output_hidden_states=True`, parse the molecule with RDKit, and align tokenizer-level atom tokens with their RDKit atom indices via a deterministic walk through the regex tokenization. Each atom-token's element symbol is verified against the corresponding RDKit atom; mismatched molecules are dropped. The per-atom vector $\mathbf{h}_{\tau,l}^{(i)} \in \mathbb{R}^d$ at layer l is the i-th atom-token row of the layer-l hidden-states tensor.

Standardization. For each (task, layer), per-feature mean $\boldsymbol{\mu}_{\tau,l}$ and standard deviation $\boldsymbol{\sigma}_{\tau,l}$ are computed on training atoms only and applied as $\tilde{\mathbf{h}} = (\mathbf{h} - \boldsymbol{\mu}_{\tau,l})/(\boldsymbol{\sigma}_{\tau,l} + 10^{-8})$ to both train and test. The 10^{-8} term avoids division by zero on near-constant features.

Reproducibility shims. The public MoLFormer release predates `transformers` ≥ 5, so we add three monkey-patches at load time: a stub for the removed `transformers.onnx` subpackage, a stub for `find_pruneable_heads_and_indices`, and a re-injection of `PreTrainedModel.get_head_mask`. We also force a recomputation of the rotary embedding's non-persistent `cos_cached`/`sin_cached` buffers after model loading and set `deterministic_eval=True` on every FAVOR+ feature map. SMI-TED's `pytorch-fast-transformers` dependency does not build under our environment; we reimplement its encoder in pure PyTorch (~ 150 lines) and load the released checkpoint with a strict assertion that only the post-encoder `LayerNorm` (which our probes do not read) is allowed to be unmatched.

C.2. Probe controls

MLP-probe control. To test whether the linear probe is bottlenecked by lack of nonlinearity, we additionally train a 2-layer MLP $\mathrm{MLPProbe}(\mathbf{x}) = \mathbf{W}_2\,\mathrm{ReLU}(\mathbf{W}_1\,\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$ with hidden size 128 on the real hidden states. Across all eight tasks and both models, the MLP probe improves test accuracy over the linear probe by at most 0.03 in absolute terms, with most improvements being ≤ 0.01. The largest gain is on MoLFormer hybridization ($+0.03$ accuracy / $+0.07$ macro-F1 at layer 5); on surface-form tasks the gap is essentially zero. We conclude that the chemical properties under study are approximately linearly accessible from the residual stream, and the conclusions drawn from the linear probe are not artifacts of probe capacity.

Random-features control. For each (task, layer) pair we additionally train an identical linear probe on a Gaussian random-features matrix $\mathbf{X}_{\mathrm{rand}} \in \mathbb{R}^{N\times d}$ with i.i.d. entries $\mathcal{N}(0,1)$, sampled once per script run and reused for every (task, layer). The same molecule-level split, class encoding, and standardization are applied; the probe sees the same atom labels but uninformative inputs. Test macro-F1 falls within 1–2 percentage points of $1/C_\tau$ (the chance level under macro-F1 when always predicting the majority class) for every task on both models, confirming that the real-features probe accuracy reflects encoder structure rather than overfitting capacity in the linear head.

C.3. Causal-steering setup

Steering vector. For each (task, steer layer ℓ) we form, in the original (unstandardized) hidden-state space,

$$\mathbf{v}^{c\to c'} = (\mathbf{W}_{c'} - \mathbf{W}_c)/\boldsymbol{\sigma},$$

where $\mathbf{W}_c$ is the layer-ℓ probe's row for class c and $\boldsymbol{\sigma}$ is the per-feature training standard deviation. Adding $\alpha \cdot \mathbf{v}^{c\to c'}$ to a hidden state increases the standardized logit difference $\mathrm{logit}_{c'} - \mathrm{logit}_c$ by $\alpha \cdot \|\mathbf{W}_{c'} - \mathbf{W}_c\|^2_{\boldsymbol{\sigma}^{-2}}$. For binary tasks, $\mathrm{tgt} = 1 - \mathrm{src}$. For multi-class tasks, we use a cycle target $\mathrm{tgt} = (\mathrm{src} + 1) \bmod C_\tau$.

Forward hook and intervention. The hook is registered on the module that produces $\mathbf{h}^{(\ell)}$ (`model.encoder.layer[`ℓ`-1]` for MoLFormer, `encoder.blocks.layers[`ℓ`-1]` for SMI-TED), modifies that layer's output by adding the per-position steering tensor at atom-token positions only, and is removed in a `try/finally` block immediately after the post-intervention forward pass. The pre- and post-intervention forward passes are run with `output_hidden_states=True` to capture the residual stream at every depth.

Layer sweep. Figure 5 sweeps $\ell \in \{1, \ldots, 12\}$ for each task and reports the intervention success rate at the best-performing α on held-out molecules. This best-α-per-layer summary serves as a single scalar diagnostic of whether the encoder routes through the layer-ℓ probe direction.

Metrics. *Intervention success rate* (or *flip-to-target rate*) is the fraction of atom-token positions whose layer-12 probe prediction equals the intended target c' after the intervention. We additionally track *moved-off-source rate* (the fraction whose post-intervention layer-12 prediction differs from the source class c) so that, for multi-class tasks, "the steering had effect" can be disentangled from "the steering hit the cycle target"—a distinction that matters when the cycle direction $\mathbf{W}_{c'} - \mathbf{W}_c$ does not point at the cycle-target centroid in the probe's prototype layout.

Sample size and hyperparameters. Each task is evaluated on 200 molecules (~ 1.7k atom-token positions; 197 molecules and $1{,}719$ atom-token positions survive token-atom alignment) drawn from the QM9 partition not used for probe training. The α grid is $\{0, 1, 3, 10, 30, 100, 300\}$; the smallest non-zero α that maximizes success rate is reported per layer.

C.4. Layer-11 α-grid sweep in SMI-TED

To rule out the possibility that the layer-11 dip in SMI-TED's intervention curves (Section 4.3) reflects a poorly chosen perturbation magnitude rather than a property of the encoder, we re-ran the layer-11 intervention on the five properties that show the dip (the four contextual properties plus formal charge) with a finer α grid in $[0, 1]$, in addition to the wide grid up to $\alpha = 300$ already used in the main sweep. Table 1 reports the per-α flip-to-target rate at the layer-12 readout. At every α in the combined grid, the success rate at layer 11 saturates well below the L10 / L12 ceiling, so the dip is not a sweep artifact.

α	is_in_ring	degree	chiral_tag	hybridization	formal_charge
0.0	0.00	0.00	0.02	0.00	0.00
0.1	0.00	0.11	0.10	0.00	0.00
0.2	0.00	0.19	0.14	0.00	0.00
0.3	0.00	0.21	0.20	0.00	0.00
0.5	0.00	0.25	0.20	0.01	0.02
0.7	0.00	0.25	0.20	0.01	0.03
1.0	0.00	0.24	0.20	0.00	0.03
3	0.00	0.13	0.21	0.01	0.03
10	0.00	0.07	0.21	0.01	0.03
30	0.00	0.06	0.21	0.01	0.03
100	0.00	0.06	0.21	0.01	0.03
300	0.00	0.06	0.21	0.01	0.03
best @ L11	**0.00**	**0.25**	**0.21**	**0.01**	**0.03**
best @ L10	1.00	1.00	1.00	0.91	0.99
best @ L12	1.00	1.00	1.00	0.59	1.00

Table 1. Intervention success rate at SMI-TED layer 11 for the five properties that show the L11 dip.