What Does a Chemical Language Model Know About Molecules?

Christian Kenneth [1] **Etowah Adams** [2] **Liam Bai** [3] **Gerard JP van Westen** [4]

Abstract

Chemical language models (cLMs) are widely assumed to learn surface-level syntactic patterns rather than learning meaningful molecular semantics. Here, we apply sparse autoencoders (SAEs) to MolFormer, an encoder-only cLM, to mechanistically examine how molecular representations are built across layers. We discover that early layers rely on position-tracking latents to parse molecular grammar, while later layers encode atom-in-substructure and pharmacologically relevant features. Additionally, we show that non-canonical SMILES produce more disruptive representation shifts than invalid SMILES, driven by position-latent disruption propagating across layers. To support further exploration, we develop InterMol, an interactive visualizer for SAE activations on molecular strings and structures.

1. Introduction

Chemical language models (cLMs, or molecular language models) have shown strong performance in molecular property prediction (Ross et al., 2022) and are widely used in generative drug design (Grisoni, 2023). cLMs learn 'chemical language' from datasets containing up to billions of molecules. Some studies have further used compound-class-specific datasets, such as natural products (Sakano et al., 2024) and polymers (Kuenneth & Ramprasad, 2023), to allow cLMs to develop specialized representations. Despite these successes, some in the community remain skeptical about whether cLMs learn meaningful chemical features or merely exploit surface-level string patterns, given that molecules can be represented as strings in many equivalent ways (Bajorath, 2024; Kikuchi et al., 2026).

Addressing these concerns requires understanding how cLMs internally represent molecules. Recent work in AI interpretability shows that sparse autoencoders (SAEs)—a method for mechanistic interpretability—can uncover human-understandable features from large language models (e.g., GPT-4 (Gao et al., 2024) and Claude Sonnet 3 (Templeton et al., 2024)) as well as domain-specific language models (e.g., ESM2 for proteins (Simon & Zou, 2025; Adams et al., 2025) and Evo2 for nucleic acids (Brixi et al., 2025)). Since cLMs share the same foundations as these language models, SAEs offer a grounded way to answer the open question of *what a chemical language model knows about molecules*.

Concurrent with our work, Cohen et al. (2025) and Varadi et al. (2025) have also applied SAEs to understand cLM representations of molecules. While these studies focus on encoder-decoder and decoder-only architectures, respectively, we train SAEs on the residual stream of an encoder-only model, MolFormer-XL, and pursue a broader investigation into what the model has learned (discussed further in Section B). To answer our central question, we advance the analysis with four main contributions that together provide new insights into cLM molecular representations:

1. **InterMol.** We developed a web application for interactive visualization of SAE latent activations, available at www.intermol.co.
2. **Molecular Grammar.** We identified SAE latents that activate on syntactic concepts, including branches, ring indices, and syntax parsing features besides atom-related features.
3. **Learned Representation Shifts.** We show that position latents make non-canonical SMILES more disruptive to learned representations than invalid SMILES.
4. **Interpretable Probing.** We identified pharmacologically relevant SAE latents from linear probe coefficients trained on downstream tasks.

2. Background

2.1. Molecular String Representations

Molecules are commonly represented as graphs, with atoms as nodes and bonds as edges. However, molecular graphs are computationally expensive for large sets of molecules.

[1]Independent [2]Department of Systems Biology, Columbia University, New York [3]Generate:Biomedicines, Massachusetts [4]Computational Drug Discovery (CDD), Division of Medicinal Chemistry, Leiden University, The Netherlands. Correspondence to: Christian Kenneth <kennethasikinnn@gmail.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

Weininger (1988) tackled this limitation by introducing SMILES, Simplified Molecular Input Line Entry System, as a compact string representation. SMILES encodes molecular graphs by traversing atoms and bonds depth-first from any starting atom. Special characters indicate specific structural features, such as parentheses for branching and digits for ring closures.

Although alternative molecular string representations have been developed, such as InChI (Heller et al., 2013), DeepSMILES (O'Boyle & Dalke, 2018), and SELFIES (Krenn et al., 2020), SMILES remains widely used due to its simplicity, human readability, widespread adoption across curated chemistry databases, and use as the input for many cLMs. Based on these advantages, we use SMILES as the main molecular string representation for our mechanistic interpretability study.

2.2. Chemical Language Models

cLMs apply natural language processing (NLP) techniques to learn from molecular string representations. Since the introduction of the transformer architecture, many cLMs have been developed to learn representations of chemical space through masked language modelling, including ChemBERTa variants (Chithrananda et al., 2020; Ahmad et al., 2022) and MolFormer-XL (Ross et al., 2022). Although ChemBERTa is widely used as a baseline, we selected MolFormer-XL for several reasons: (1) rotational position embeddings that allow efficient processing of long molecular strings; (2) pretraining on a large and diverse molecular corpus of 1.1 billion compounds from ZINC15 and PubChem compared to 77 million compounds from PubChem used for ChemBERTa; and (3) superior performance on most chemistry-related benchmarks.

In particular, we investigation the open-weights variant of MolFormer-XL (Ross et al., 2022). This variant was pretrained on only 10% of the molecular SMILES used for the full model, yet achieved comparable performance on several chemistry benchmarks. Hereafter, we refer to this variant as MolFormer-10, which serves as the main model for our interpretability analysis.

3. Methods

3.1. Sparse Autoencoders

Sparse autoencoders (SAEs) are trained to reconstruct inputs while enforcing sparsity in the hidden layer. In this study, we use k-SAE (Gao et al., 2024), which keeps only the top k hidden-layer activations z and nullifies the rest. The computation is expressed as:

$$z = \text{TopK}(W_{enc}x + b_{enc})$$

$$\hat{x} = W_{dec}z + b_{dec}$$

where the encoder weights W_{enc} project the residual stream into the SAE latent space and the decoder weights W_{dec} reverse it. The reconstruction mean squared error $\mathcal{L} = \|x - \hat{x}\|_2^2$ is used as the objective function to train the SAE.

3.2. Molecular Features

As part of understanding what the cLMs have learned from SMILES, we classified each SAE latent by its dominant token types and further map SAE activations to atom-related concepts: atom, atom-in-substructure, and substructure. For these concepts, we used SMARTS (SMILES Arbitrary Target Specification) notation, as SMARTS allows us to define specific chemical patterns that are easily interpretable. We provide the detailed steps in Section C.2.2 (token types) and Section C.2.3 (atom-related concepts).

3.3. SMILES Variants

Non-canonical SMILES are variations that result from generating SMILES strings because molecules as graphs can be parsed at any starting atom or traversed in different ways. This can pose challenges for cLMs pretrained only on canonical SMILES, including MolFormer-10. To assess how MolFormer-10 represents such variation, we built a toy dataset using RDKit (Landrum et al., 2025).

Invalid SMILES can appear as artifacts in generative model outputs or be accidentally introduced in databases (e.g., mistakenly marking aromatic carbons `c` as non-aromatic ones `C`). As Skinnider (2024) suggests that invalid SMILES serve a beneficial role in generative models, understanding how cLMs internally represent them could inform the design of future cLMs. To this end, we adapted SMILES-corrupting scripts from Schoenmaker et al. (2023) to build the toy dataset with further details in Table 1.

We use RDKit-canonicalized SMILES throughout this paper unless otherwise specified.

3.4. Downstream Tasks

ADMET, or Absorption, Distribution, Metabolism, Excretion, and Toxicity, describes how drugs are processed by the body and is critical in ensuring that drugs are both safe and effective. We used ADMET prediction tasks, comprising 11 classification (Table 2) and 10 regression tasks (Table 3), to evaluate whether cLMs learned pharmacologically relevant features. The datasets were retrieved from the Therapeutics Data Commons (TDC) (Huang et al., 2021; 2022).

4. Decomposing Molecular Understanding

Uncovering chemically meaningful features from SAE latents is challenging due to the diverse token types in SMILES. SMILES consists of five main token types: atoms,

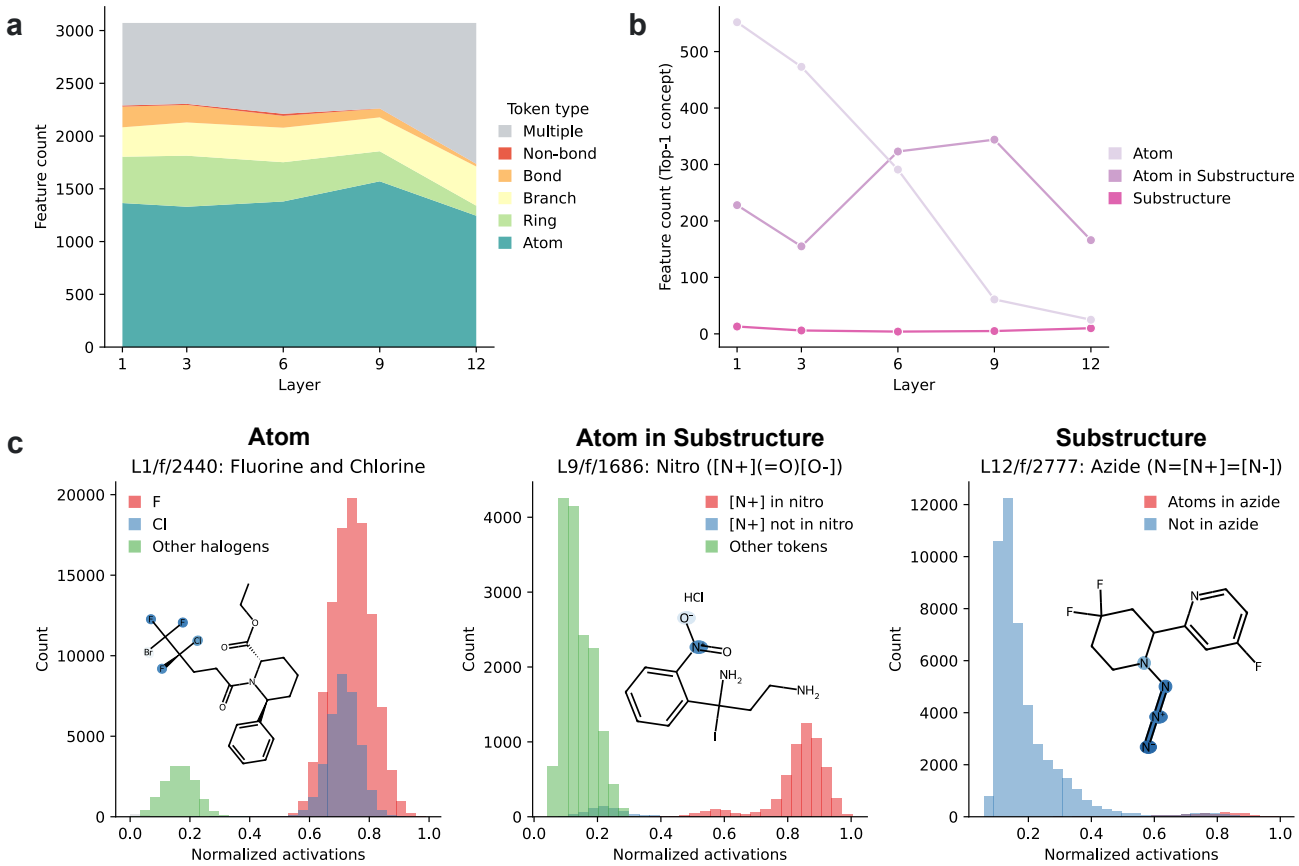

Figure 1. **Interpretable Molecular Features. (a)** SAE latents classified by dominant token type ($D \geq 0.6$) across layers. **(b)** Number of unique interpretable latents per layer matching their top-1 atom-related concept. **(c)** Examples of atom-related concepts encoded in latents across layers.

ring indices that number rings, branches marking branching points, bond notations representing interatomic connectivity (e.g., = for double bonds and # for triple bonds), and disconnections or non-bonds denoting molecular fragments.

Examining the token-type distribution of each latent across MolFormer-10 layers via Simpson's dominance index D reveals that syntax-related token types are least frequent across layers, while nearly half of the latents up to Layer 9 activate on atom-type tokens (Figure 1a). This is because syntax-related tokens—rings, branches, bonds, and non-bonds—contain fewer unique tokens by their nature as molecular grammar rather than atomic composition. In the last layer, however, latents activating on multiple token types become as common as those activating on atom types, signifying more diffuse activation patterns compared to the token-specific patterns observed in early and middle layers.

4.1. Atom-Related Concepts

We next match atom-type-preferring latents against a predefined library of atom-related concepts (Section C.2.3). By selecting the concept with the highest F1 score (F1 ≥ 0.5) for each latent, we measure the number of interpretable latents in Figure 1b, with several examples shown in Figure 1c.

Across layers, latents matching atom-level concepts show a bottom-up trend, from atom-focused activation to atom-in-context and substructure activation patterns (Figure 1b). Atom-in-context latents suggest a more efficient SAE approach to decomposing molecules based on common structural contexts, especially in Layers 6 and 9. For example, rather than learning separate latents for carboxylic acids (`C(=O)O`) and amides (`C(=O)N`), the model appears to learn shared atom-pair latents for their common motifs: `C=O`, `C-N`, and `C-O`. When an amide is present, the first two latents activate together; when a carboxylic acid is present, the first and third activate. This decomposition may enable the cLM to reuse these latents across more complex molecular structures.

Substructure-level concepts were expected to enrich in later layers, but instead show an unclear trend as seen in Figure 1b. Nonetheless, given that most of these latents activate across multiple token types (Figure 1a), we argue that this

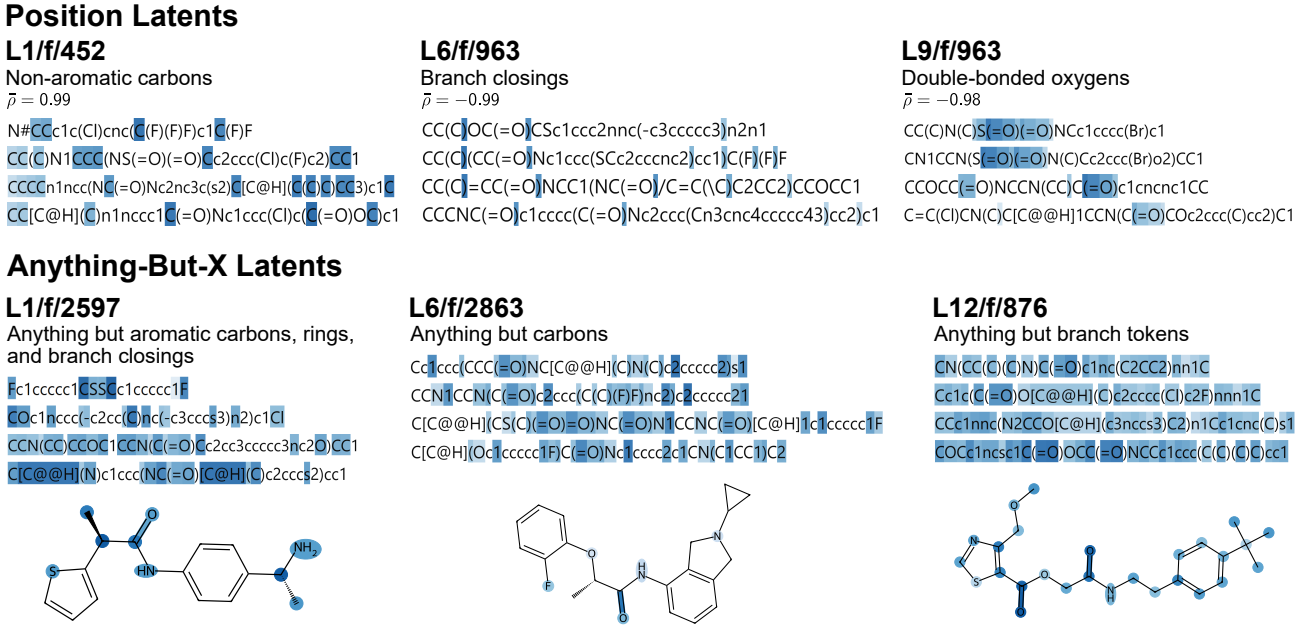

Figure 2. **Syntax Parsing Features. Position Latents**: **L1/f/452** shows increasing activations on token `C` toward the final token; **L6/f/963** shows decreasing activations on branch-closing token `)` toward the end; and **L9/f/1494** shows decreasing activations on the `(=O)` motif, possibly tracking double-bonded oxygens. **Anything-But-X Latents**: **L1/f/2597** ($p_{\text{zero}} = 0.99$) does not activate on aromatic carbons ($p_{\text{ctx}} = 1.0$), ring indices ($p_{\text{ctx}} = 0.99$), and branch closings ($p_{\text{ctx}} = 0.99$); **L6/f/2863** ($p_{\text{zero}} = 0.99$) does not activate on any carbons, including `C` ($p_{\text{ctx}} = 0.99$), `c` ($p_{\text{ctx}} = 0.99$), `[C@H]` ($p_{\text{ctx}} = 0.99$), `[C@@H]` ($p_{\text{ctx}} = 0.99$), `[C@]` ($p_{\text{ctx}} = 0.95$), and `[C@@]` ($p_{\text{ctx}} = 0.99$); and **L12/f/876** ($p_{\text{zero}} = 0.98$) does not activate on branch openings ($p_{\text{ctx}} = 0.99$) and closings ($p_{\text{ctx}} = 0.99$).

indicates the cLM may not learn all structural contexts, but rather syntactic ones of SMILES.

4.2. Syntax Parsing Features

We also discover latents whose activation patterns extend beyond atom-related concepts. We term these syntax parsing features, arguing that they may provide contextual information enabling the cLM to process inputs and build internal molecular representations. We identify two types: position latents and anything-but-X latents.

4.2.1. Position Latents

A recent study (Sun et al., 2025) found that several dense latents (i.e., those activating on multiple tokens) have activations correlated with relative token positions within specific contexts. We identify similar latents in our SAE, which are more prominent in early layers and decrease toward later layers Figure 4, consistent with early layers processing the tokenized SMILES input. As shown in Figure 2, L1/f/452 and L6/f/963 strongly activate on later non-aromatic carbon tokens and earlier branch-closing tokens, respectively, and decrease toward the opposite end with no preference of specific context. Unlike simple position tracking, L9/f/1494 tracks double-bonded oxygens within a branch from opening to closing, indicating syntactic substructure completeness.

4.2.2. Anything-But-X Latents

Anything-but-X latents activate on all tokens except specific tokens or contexts. Previously, Fel et al. (2025) identified similar latents, termed "elsewhere" latents, in a vision transformer. There are two points that make anything-but-X latents differ from "elsewhere" latents: (1) "elsewhere" latents activation is context-dependent, whereas ours activate independently; and (2) consequently, "elsewhere" latents imply learned context negations, while the function of anything-but-X latents remains unclear—possibly an artifact of the small SAE size.

We characterize these latents using two proportions that measure latents exclusivity: (1) the ratio of non-activated tokens within a specific context to all **non-activated tokens** (p_{zero}), and (2) the ratio of non-activated tokens within a specific context to all **tokens with that context in the dataset** (p_{ctx}). As an example (Figure 2), L6/f/2863 exhibits this characteristic by not activating on common carbon-related tokens, suggesting it may help discriminate heteroatoms. While this analysis serves as a preliminary finding, a more thorough analysis using larger SAE dimensions and circuit-level analysis of the cLM residual stream is needed to confirm the exclusivity of anything-but-X latents and rule out polysemanticity.

5. SMILES Variants Perturb Representations

CLMs that learn molecular structures should ideally be invariant to variations in molecular string notation. To test this, we analyze how cLM internal representations encode SMILES variants by examining the similarity of max-pooled SAE latents (SAE embeddings) and assessing how individual latents respond to shifts in internal molecular representation.

For SAE representation similarity, we compute cosine similarity between raw values and Jaccard similarity between binarized (active/inactive) values across canonical and variant SMILES pairs. For the latent-specific analysis, we compute the standardized mean difference (SMD) as an effect size for each latent f.

$$\text{SMD}_f = \frac{\mu_f^c - \mu_f^{nc}}{\sigma_{\text{pooled}}}$$

where μ_f^c and μ_f^{nc} are the mean max-pooled activations of latent f over the reference (canonical or valid) and comparison (non-canonical or invalid) groups, respectively, and σ_{pooled} is the pooled standard deviation across both groups. Results of both analyses for the augmented SMILES are presented in Figure 4a and Figure 4b, respectively.

5.1. Augmented SMILES

As illustrated in Figure 4a, the median cosine similarity of SAE embeddings between canonical and non-canonical SMILES is higher than the median Jaccard similarity across layers, with the largest discrepancy at layer 12. This indicates that while the set of active latents may differ between canonical and non-canonical SMILES, those that activate in both share similar activation magnitudes, suggesting they likely capture structural rather than purely syntactic patterns.

Latents with large effect size ($|\text{SMD}| \geq 0.8$) are mostly position latents in early layers (Figure 4b). For example, as shown in Figure 4c, non-canonical SMILES suppress the max-pooled activations of L1/f/1627 when the first token changes from `C` to `N`. From layer 6 onward, non-position latents with large $|\text{SMD}|$ that activate more strongly in non-canonical SMILES predominate (Figure 8), suggesting that early-layer position latents disruption may propagate into deeper layers.

5.2. Corrupted SMILES

Of the five error types, syntax errors exhibit the widest similarities distribution (Figure 7), likely because they introduce more disruptive modifications, such as SMILES truncation, compared to simpler token insertion or duplication in other error types (see Table 1). Interestingly, aromaticity and kekulization errors have the highest median similarities across layers, despite mainly swapping atoms that contribute differently to the ring's π-electron system. This suggests that the SAE recognizes whether an atom belongs to an aromatic ring, but is insensitive to the specific atomic identity within that ring.

Regarding latent-specific differences, almost all error types affect few latents significantly, except for valence errors (Figure 9). This is because valence errors introduce new bond tokens that activate related latents, resulting in high-SMD latents that are primarily non-positional. Notably, we identified a latent in layer 12, L12/f/956, that consistently shows a large effect size across all SMILES variants, with stronger activation when the input is augmented (Figure 8) or corrupted (Figure 9), implying that this latent may function as a canonicality and validity detector.

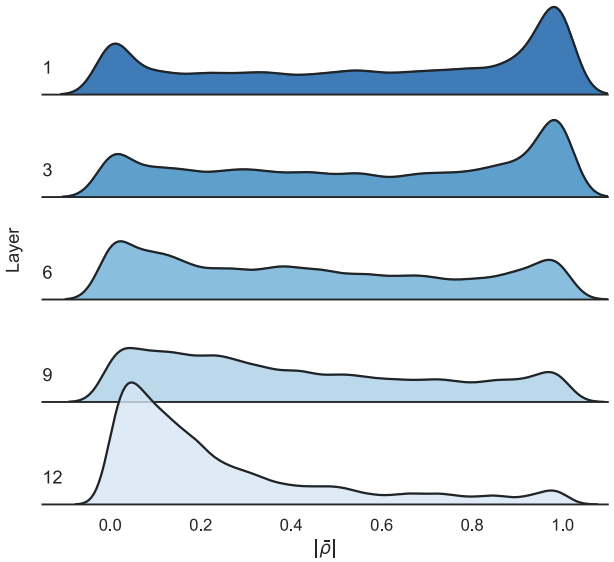

Figure 4. **Position Latents Decrease Toward Later Layers.** Distribution of absolute Fisher-averaged Spearman coefficients across layers. Peaks near $|\bar{\rho}| \approx 1$ at early layers indicate a significant presence of position latents, while later layers show a flattened distribution skewed toward $|\bar{\rho}| \approx 0$, suggesting that fewer position latents are present and contribute less positional information to the layer's internal representation.

6. Chemically Relevant Latents

We used ADMET tasks to search for chemically relevant latents. Following Adams et al. (2025), we employ SAE latents as input features to a simple linear model, whose coefficients directly reflect the contribution of each latent. The InterMol visualizer is used to interpret the top-contributing latents and identify their physicochemical-related features.

6.1. Performance on ADMET Tasks

Overall, SAE embeddings perform better on 16 out of 21 ADMET tasks, followed by MolFormer-10 (4 tasks), and

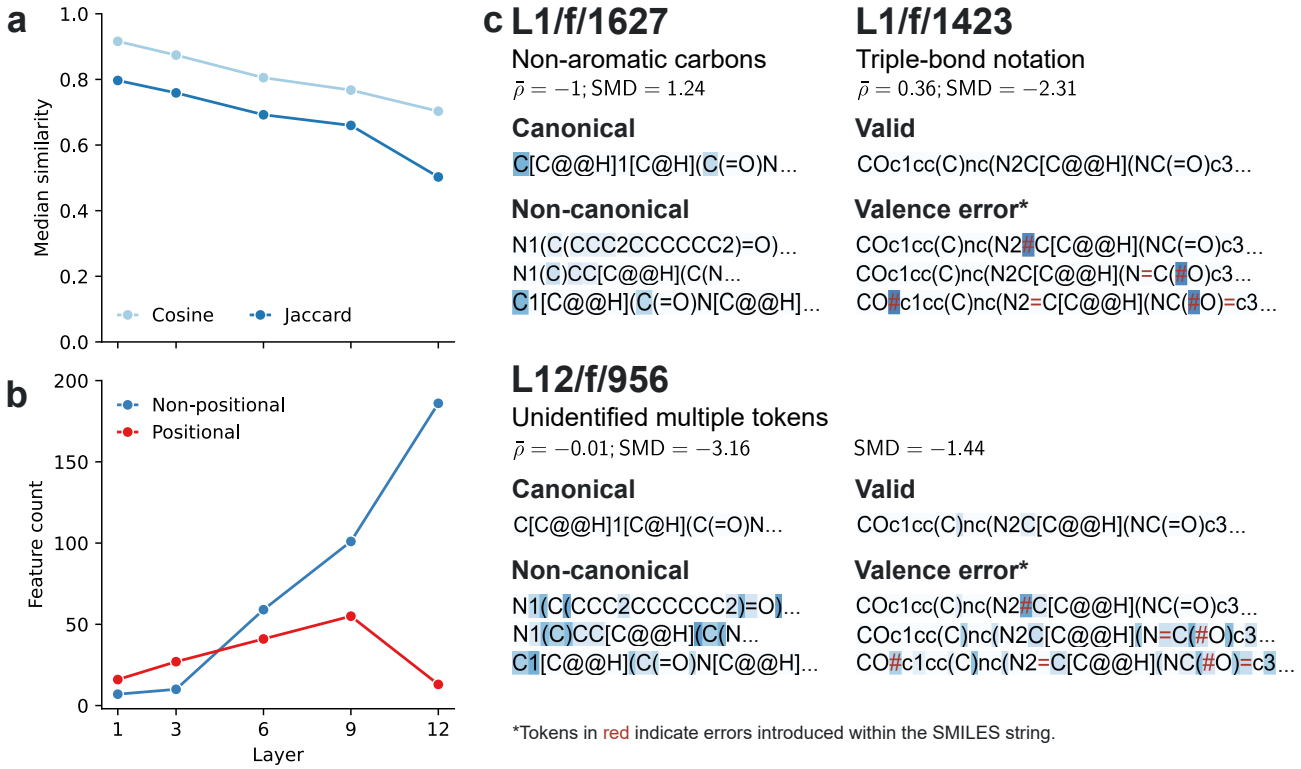

Figure 3. **Position Latents Shift Learned Representations.** **(a)** Median cosine and Jaccard similarities between max-pooled SAE activations of canonical and non-canonical SMILES pairs. **(b)** Number of SAE latents with $|\text{SMD}| \geq 0.8$ classified as non-positional and positional across layers for the augmented SMILES. **(c)** Examples of features with significant SMD in non-canonical and valence error cases.

physicochemical descriptors (1 task). SAE latents likely capture chemical features similar to the baselines, such as atom-in-substructure concepts analogous to the atomic environments encoded in ECFPs (Figure 12), while potentially encoding more complex features not explicitly captured by the baselines. We examine 4 tasks for further pharmacological interpretation in Section 6.2.

6.2. Interpreting Probes

Human Intestinal Absorption (HIA) evaluates the ability of drug molecules to be absorbed into the bloodstream from the gastrointestinal (GI) tract (Hou et al., 2007). Our linear probe identifies three latents corresponding to two structural motifs associated with absorptivity (Figure 6a). The first is halogen substituents, such as chloro- (L9/f/2608) and fluoro- (L9/f/2705), both of which are known to improve lipophilicity and thus GI absorptivity (Dudek et al., 2025). The second is double-bonded oxygen; L9/f/1499 specifically fires on the double bond token. Although this motif increases solubility, a balance between solubility and lipophilicity is needed for optimal GI absorption—making this latent meaningful in predicting HIA.

CYP2C9 Inhibition. CYP2C9 is a cytochrome P450 isoenzyme whose inhibition hinders the body's ability to metabolize co-administered drugs, potentially leading to toxicity (Veith et al., 2009). Among the three SAE latents identified, only one is pharmacologically relevant: L12/f/1735, which activates on heteroatoms in heterocycles, aligned with common pharmacophoric features of CYP2C9 inhibitors reported in Beck et al. (2021). However, other pharmacophoric motifs absent from the top-3 latents, particularly relevant for inhibitory tasks, may reside in other SAE latents as supported by Figure 11. While L12/f/1750 has no clear interpretation, its activation on the oxygen atom between aromatic rings (Figure 6b) is able to differentiate between two molecules with similar structures but different inhibitory activity.

Ames Mutagenicity. This refers to the ability of a molecule to induce DNA damage and cause genetic alterations. The dataset is constructed from the Ames bacterial test, the most widely used *in vitro* assay for assessing mutagenic potential (Xu et al., 2012). We found three latents corresponding to mutagenic properties. The top latent, L9/f/12, activates on nitrogens bonded to other nitrogens, commonly found in mutagenic motifs such as the N-N=O of nitrosamines. The remaining two latents both activate more strongly on `[N+]`: L9/f/1686 on nitro groups and L9/f/2807 on azides

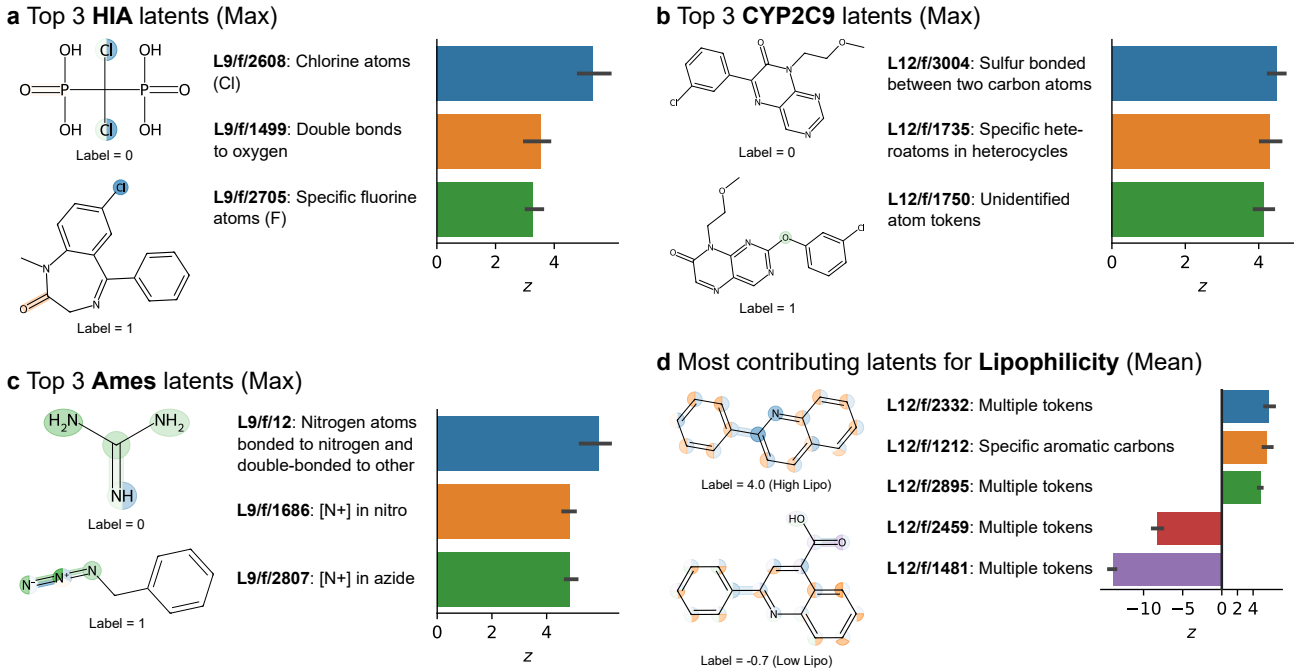

Figure 5. **Linear Probe Identifies Pharmacologically Relevant SAE Latents.** We show the top 3 latents by standardized linear probe coefficient, $z(\beta_f)$, averaged across triplicates, for **(a)** human intestinal absorption (HIA), **(b)** CYP2C9 inhibition, and **(c)** Ames mutagenicity classification; and **(d)** top 3 positive and top 2 negative latents for the lipophilicity regression task.

(the latter seen in Figure 6c), both well-known mutagenic motifs (Nepali et al., 2019; Grúz et al., 2022).

Lipophilicity measures a molecule's relative tendency to dissolve in lipids over water, directly influencing its ability to be absorbed in the GI tract. As shown in Figure 6d, no latents are found to activate on specific concepts; however, the activation of L12/f/1212 on aromatic carbons provides an interpretable basis for increased lipophilicity, as aromatic carbons are known to be non-polar. In contrast, for the low lipophilicity sample, the additional functional motif -COOH is activated with L12/f/1481, which negatively influences the predicted lipophilicity, consistent with the lipophobic nature of the motif.

7. Discussion

In this work, we presented a mechanistic analysis of an encoder-only cLM, MolFormer, using SAEs. We found that molecular representations are built hierarchically, progressing from atom-level concepts at early layers toward more nuanced atom-in-substructure and substructure-level concepts at later layers. This hierarchical organisation is further reflected in the performance on ADMET tasks, where most results exceed the baseline methods of molecular fingerprints and descriptors, showing that MolFormer encodes pharmacologically relevant molecular features. To support the further exploration of many remaining uninterpreted latents, we introduce InterMol, an open-source visualizer for that may reveal new insights into the inner workings of cLMs.

In addition to atom-related features, MolFormer also has its own mechanistic way of parsing the molecular grammar encoded in SMILES input, using prevalent position latents to track atom-type token positions and syntactic substructure completeness. Yet, this mechanism may introduce an invariance challenge for cLMs pretrained only on canonical SMILES, as it may cause position-dependent disruptions to learned representations across different SMILES notations of the same molecule. Therefore, a good training design, like exposing the cLM to augmented SMILES or, with incorporating contrastive learning, may help address this issue by reducing positional bias, as supported by recent studies (Bjerrum, 2017; Brinkmann et al., 2025). On the contrary, invalid SMILES produce more subtle shifts in learned representations. This may also partially explain the behavior of autoregressive cLMs that is occasionally generate invalid SMILES during molecular generation, though additional study is needed to confirm this.

Limitations. Despite these findings, our work has several limitations. First, we focus only on MolFormer, which may not generalize to other encoder-only cLMs due to differences in architecture, SMILES input, or atom-wise tokenization, though our findings provide a foundation for future cLM designs, with MolFormer itself already aligned with recent recommendations of (Fender et al., 2025). Sec-

ond, although we randomly sampled from available molecular SMILES to reduce structural biases, our SAE training dataset remains limited in scale. Third, MolFormer uses rotary position embeddings (RoPE) in each attention head computation, which entangles relative positional information and molecular language signals, making attention-based circuit analysis difficult to apply. We therefore limit our analysis to the molecular features encoded in the residual stream of MolFormer. Finally, interpreting SAE latents remains challenging given the lack of gold-standard molecular concepts, unlike those available for proteins via UniProt and DNA via GenBank, and we therefore consider some of our interpretations of complex activation patterns as preliminary—we invite the community to refine or correct them through InterMol.

Future Work. Our findings suggest several promising directions for future research. One direction involves improving the training design of cLMs. While SMILES augmentation may help reduce positional bias as previously discussed, multimodal representation learning offers a more grounded approach. For instance, joint-embedding predictive architecture (JEPA) could be incorporated by integrating both molecular strings and structures, following recent evidence of successful LLM-JEPA integration (Huang et al., 2025). This multimodality may help cLMs learn richer molecular representations, less dominated by syntax-related features. Another intriguing direction is deepening the mechanistic analysis of cLMs by employing cross-layer transcoders (CLTs) to identify more causally grounded molecular cirucits, recently applied to understand protein circuits in protein language models (Tsui et al., 2026) and enable downstream applications like steering autoregressive cLMs for self-directed molecular generation.

Acknowledgements

We declare no conflicts of interest.

Impact Statement

This work advances our understanding of cLMs by revealing interpretable molecular features encoded in their representations. Our findings may support rational molecular design and accelerate drug discovery. We do not foresee ethical risks arising from this research.

References

Adams, E., Bai, L., Lee, M., Yu, Y., and AlQuraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025. doi: 10.1101/2025.02.06.636901. URL https://www.biorxiv.org/content/early/2025/06/18/2025.02.06.636901.

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta-2: Towards chemical foundation models, 2022. URL https://arxiv.org/abs/2209.01712.

Bajorath, J. Chemical and biological language models in molecular design: opportunities, risks and scientific reasoning. *Future Sci. OA*, 10(1):FSO957, May 2024.

Beck, T. C., Beck, K. R., Morningstar, J., Benjamin, M. M., and Norris, R. A. Descriptors of cytochrome inhibitors and useful machine learning based methods for the design of safer drugs. *Pharmaceuticals (Basel)*, 14(5):472, May 2021.

Bjerrum, E. J. Smiles enumeration as data augmentation for neural network modeling of molecules, 2017. URL https://arxiv.org/abs/1703.07076.

Brinkmann, H., Argante, A., ter Steege, H., and Grisoni, F. Going beyond smiles enumeration for data augmentation in generative drug discovery. *Digital Discovery*, 4:2752–2764, 2025. doi: 10.1039/D5DD00028A. URL http://dx.doi.org/10.1039/D5DD00028A.

Brixi, G., Durrant, M. G., Ku, J., Poli, M., Brockman, G., Chang, D., Gonzalez, G. A., King, S. H., Li, D. B., Merchant, A. T., Naghipourfar, M., Nguyen, E., Ricci-Tam, C., Romero, D. W., Sun, G., Taghibakshi, A., Vorontsov, A., Yang, B., Deng, M., Gorton, L., Nguyen, N., Wang, N. K., Adams, E., Baccus, S. A., Dillmann, S., Ermon, S., Guo, D., Ilango, R., Janik, K., Lu, A. X., Mehta, R., Mofrad, M. R., Ng, M. Y., Pannu, J., Ré, C., Schmok, J. C., John, J. S., Sullivan, J., Zhu, K., Zynda, G., Balsam, D., Collison, P., Costa, A. B., Hernandez-Boussard, T., Ho, E., Liu, M.-Y., McGrath, T., Powell, K., Burke, D. P., Goodarzi, H., Hsu, P. D., and Hie, B. L. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025. doi: 10.1101/2025.02.18.638918. URL https://www.biorxiv.org/content/early/2025/02/21/2025.02.18.638918.

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.

Cohen, J., Hasson, A. G., and Tanovic, S. Unveiling latent knowledge in chemistry language models through sparse autoencoders, 2025. URL https://arxiv.org/abs/2512.08077.

Dudek, A., Dejnaka, E., Sulecka-Zadka, J., Perz, M., Krawczyk-Łebek, A., Kostrzewa-Susłow, E., Pruchnik,

H., and Pawlak, A. Bromo- and chloro-substituted flavones induce apoptosis and modulate cell death pathways in canine lymphoma and leukemia cells - a comparative in vitro study. *Front. Mol. Biosci.*, 12(1738255): 1738255, 2025.

Fel, T., Wang, B., Lepori, M. A., Kowal, M., Lee, A., Balestriero, R., Joseph, S., Lubana, E. S., Konkle, T., Ba, D., and Wattenberg, M. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry, 2025. URL https://arxiv.org/abs/2510.08638.

Fender, I., Gut, J. A., and Lemmin, T. Beyond performance: how design choices shape chemical language models. *J. Cheminform.*, 17(1):173, November 2025.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Grisoni, F. Chemical language models for de novo drug design: Challenges and opportunities. *Curr. Opin. Struct. Biol.*, 79(102527):102527, April 2023.

Grúz, P., Yasui, M., Ukai, A., Horibata, K., Honma, M., and Sugiyama, K.-I. Potent mutagenicity of an azide, 3-azido-1,2-propanediol, in human TK6 cells. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.*, 876-877(503475):503475, April 2022.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.*, 5(1):7, January 2013.

Hou, T., Wang, J., Zhang, W., and Xu, X. ADME evaluation in drug discovery. 7. prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.*, 47 (1):208–218, January 2007.

Huang, H., LeCun, Y., and Balestriero, R. Llm-jepa: Large language models meet joint embedding predictive architectures, 2025. URL https://arxiv.org/abs/2509.14252.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development, 2021. URL https://arxiv.org/abs/2102.09548.

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.*, 18(10):1033–1036, October 2022.

Irwin, J. J. and Shoichet, B. K. ZINC–a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, 45(1):177–182, January 2005.

Kikuchi, Y., Yoshikai, Y., Nemoto, S., Furuhama, A., Yamada, T., Kusuhara, H., and Mizuno, T. Notation-level confounding: When inconsistent molecular notations mislead chemical language models, 2026. URL https://arxiv.org/abs/2505.07139.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, 47(D1):D1102–D1109, January 2019.

Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1(4):045024, December 2020.

Kuenneth, C. and Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.*, 14(1):4099, July 2023.

Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., sriniker, Gedeck, P., Jones, G., Kawashima, E., NadineSchneider, Nealschneider, D., Dalke, A., tadhurst-cdd, Swain, M., Cole, B., Turk, S., Savelev, A., Maeder, N., Vaucher, A., Wójcikowski, M., Faara, H., Take, I., Walker, R., Scalfani, V. F., Probst, D., Ujihara, K., Pahl, A., Godin, G., and Lehtivarjo, J. rdkit/rdkit: 2025_09_4 (q3 2025) release, 2025.

Nepali, K., Lee, H.-Y., and Liou, J.-P. Nitro-group-containing drugs. *J. Med. Chem.*, 62(6):2851–2893, March 2019.

O'Boyle, N. and Dalke, A. DeepSMILES: An adaptation of SMILES for use in machine-learning of chemical structures, September 2018.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mach. Intell.*, 4(12):1256–1264, December 2022.

Sakano, K., Furui, K., and Ohue, M. Npgpt: Natural product-like compound generation with gpt-based chemical language models, 2024. URL https://arxiv.org/abs/2411.12886.

Schoenmaker, L., Béquignon, O. J. M., Jespers, W., and van Westen, G. J. P. UnCorrupt SMILES: a novel approach to de novo design. *J. Cheminform.*, 15(1):22, February 2023.

Simon, E. and Zou, J. InterPLM: discovering interpretable features in protein language models via sparse autoencoders. *Nat. Methods*, 22(10):2107–2117, October 2025.

Skinnider, M. A. Invalid SMILES are beneficial rather than detrimental to chemical language models. *Nat. Mach. Intell.*, 6(4):437–448, March 2024.

Sun, X., Stolfo, A., Engels, J., Wu, B., Rajamanoharan, S., Sachan, M., and Tegmark, M. Dense sae latents are features, not bugs, 2025. URL https://arxiv.org/abs/2506.15679.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tsui, D., Talreja, K., Saeedi, D., and Aghazadeh, A. Protein circuit tracing via cross-layer transcoders, 2026. URL https://arxiv.org/abs/2602.12026.

Varadi, K., Marosi, M., and Antal, P. Circuits, features, and heuristics in molecular transformers, 2025. URL https://arxiv.org/abs/2512.09757.

Veith, H., Southall, N., Huang, R., James, T., Fayne, D., Artemenko, N., Shen, M., Inglese, J., Austin, C. P., Lloyd, D. G., and Auld, D. S. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nat. Biotechnol.*, 27(11):1050–1055, November 2009.

Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, February 1988.

Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., and Tang, Y. In silico prediction of chemical ames mutagenicity. *J. Chem. Inf. Model.*, 52(11):2840–2847, November 2012.

Appendix

A. Code Availability

The code to reproduce the experiments is available at `https://github.com/ckennetha/intermol`.

B. Related Work

Mechanistic Interpretability of Chemical Language Models. Understanding how cLMs work internally may help in designing better cLMs that enable the generation of *de novo* molecules with desired properties. In parallel with our work, Cohen et al. (2025) and Varadi et al. (2025) have explored this by applying SAEs to an encoder-decoder model (SMI-TED) and and a decoder-only model, respectively. Both studies find that SAE latents correspond to chemically meaningful concepts, including substructure motifs and physicochemical properties. Cohen et al. (2025) also demonstrates pharmacological-level abstractions through features that group structurally dissimilar compounds sharing certain functionality. Both works also show controllable generation of specifc molecules through SAE-based interventions. Additionally, Varadi et al. (2025) extend their analysis to circuit-level mechanisms, identifying attention heads that track ring and branch completeness and a linear representation of valence capacity in the residual stream that modulates bond-order predictions. Despite these contributions, existing studies have not yet examined encoder-only models, which are widely used for predictive tasks. Furthermore, the global role of specific latents in capturing syntactic variation across SMILES representations has not yet been elucidated, which is integral to the development of cLMs.

C. Experimental Details

C.1. Sparse Autoencoders

Training Setup. We trained the SAEs on the 768-dimensional residual streams of MolFormer-10 layers: 1 (early), 3, 6, 9, and 12 (final), with the SAE latent dimension of 3,072 (4x expansion factor) and $k = 128$. For the training dataset, we used approximately 2.2 million SMILES randomly sampled from ZINC15 (Irwin & Shoichet, 2005) and PubChem (Kim et al., 2019), the same data pool used to train MolFormer-XL.

Normalization. For ease of interpretation, we normalized SAE latents using ∼250,000 molecular SMILES randomly sampled from the SAE training dataset. Following Simon & Zou (2025), for each latent in each SAE, we used its maximum activation to rescale the encoder and decoder weights, resulting the activations to the range [0, 1] while preserving reconstruction. However, there might be some edge cases in which the activations exceed 1.

C.2. Feature Discovery in the cLM

C.2.1. Datasets

The dataset for the feature analysis was constructed by randomly sampling another 500,000 SMILES from the same data pool that are non-overlapping with the training dataset. We refer to this as evaluation set. Then, we randomly split the evaluation set in half into validation and test sets.

In SMILES variation experiments, we randomly picked 20,000 SMILES from the validation set. For the augmented toy dataset, we generated up to 5 non-canonical SMILES per canonical SMILES, resulting in around 100,000 canonical/non-canonical pairs. For the perturbed toy dataset, we generated all possible combinations of error counts (1, 2, 4) and five error types: syntax, parentheses (unmatched barckets), rings (unclosed rings), aromaticity (non-ring atoms marked as aromatic and kekulization errors), and valence (atoms exceeding the maximum number of bonds), as shown in Table 1. This yields around 300,000 valid/invalid pairs.

Table 1. **Invalid SMILES Generation.** Error variations are randomly introduced for each error type with examples of resulting invalid SMILES. Red tokens on the left indicate deletions, while those on the right indicate substitutions or insertions.

Error type	Variation	Invalid SMILES
Rings	Remove a ring index	`c1ccccc1` → `cccccc1`
	Replace a ring index by ±1	`c1ccccc1` → `c2ccccc1`

Error type	Variation	Invalid SMILES
	Replace a ring index with different existing ring index	`C1C2CC12` → `C2C2CC12`
	Duplicate a ring opening index	`c1ccccc1` → `c11ccccc1`
	Replace a random token with a ring index	`c1ccccc1` → `c1cc1cc1`
Parentheses	Insert a random `(` or `)` token	`CC(C)O` → `C(C(C)O`
	Remove a random branch token	`CC(C)O` → `CCC)O`
	Switch a pair of parentheses	`CC(C)O` → `CC)C(O`
	Switch a branch opening token `(` to `)`	`CC(C)O` → `CC)C)O`
	Switch a branch closing token `)` to `(`	`CC(C)O` → `CC(C(O`
Aromaticity	Convert a non-aromatic to aromatic	`c1ccoc1` → `c1ccoC1`
	Convert an aromatic atom with π [`c`, `n`] to 2π electron	`c1ccoc1` → `c1ccos1`
	Insert a random `c` or `n` token into aromatic ring	`c1ccoc1` → `c1cconc1`
	Convert an aromatic atom with 2π [`[nH]`, `o`, `s`] to π electron	`c1ccoc1` → `c1cccc1`
	Remove an aromatic atom with π electron	`c1ccoc1` → `c1cco1`
Syntax	Insert a random bond [`-`, `=`, `#`] or branch [`(`, `)`] token at start	`CC(=O)CO` → `=CC(=O)CO`
	Insert a random bond or branch token at end	`CC(=O)CO` → `CC(=O)CO(`
	Insert a random bond next to an existing bond token	`CC(=O)CO` → `CC(#=O)CO`
	Insert a random bond token before a ring opening	`c1ccccc1` → `c=1ccccc1`
	Insert a random bond token before branch opening	`CC(=O)CO` → `CC-(=O)CO`
	Duplicate a branch opening token	`CC(=O)CO` → `CC((=O)CO`
	Remove tokens between `(` and `)` and optionally insert a random bond or branch token	`CC(=O)CO` → `CC()CO`
Valence	Increase the bond order of 1 to 2 or 3 and 2 to 3	`CC(=O)CO` → `C=C(=O)CO`

C.2.2. Dominant Token Type Classification

MolFormer tokenizes SMILES input using atom-wise tokenization (Ross et al., 2022), which produces tokens that can be classified further into separate token types. To characterize which token types each latent preferentially activates on, we gather all activated tokens across the validation set and grouped them by type using regex pattern matching. Tokens matching `[\(\)]` are classified as branch type, `[\-\=\#\$\\\/]` as bond type, `(\%[0-9]{2}|[0-9])` as ring type, `.` as non-bond type, and the rest as atom type.

Metric. We use Simpson's dominance index D, which is the complement of Simpson's diversity index $1 - D$, using the following formula.

$$D = 1 - \sum_{i=1}^{n_t} p_i^2$$

where p_i is the proportion of token type i across n_t total unique token types for a given latent f. Then, we assign the token type with the highest proportion as the dominant token type if $D \geq 0.6$; otherwise, we label it as "Multiple", indicating it activates on multiple token types.

C.2.3. Atom-Related Concept Association

Atom-related concepts are defined using SMARTS patterns from two sources: predefined SMARTS libraries from the RDKit GitHub repository and predominant atom-in-substructure SMARTS discovered in the validation set. The former comprises 533 patterns related to common functional motifs, which were later expanded to 2,473 unique patterns by re-enumerating each atom within every substructure. The latter were generated for each atom by incorporating its neighboring atoms and connectivity, deduplicated for the same-substructure SMARTS, and selected with occurrence in more than 10 molecules, totaling 2,483 unique patterns. Each of these patterns was then grouped into three categories: **atom** (single atom; e.g., `Cl` for chlorine atoms), **atom-in-substructure** (single atom within a substructure; e.g., `[$([N+])](=O)[O-]` for cationic nitrogen in nitro group), and **substructure** (group of connected atoms; e.g., `N=[N+]=[N-]` for azides).

Filtering SAE Latents. Before evaluating the association between atom-related concepts (i.e., pattern) and SAE latents, we employed latent filtering by selecting the top 64 latents using the standardized mean difference (SMD) between the activations on atom tokens on which the pattern is present, wc, and the activations on atom tokens on which the pattern is

not present, oc, computed on 20% of the validation set. For each latent f, we compute:

$$\text{SMD}_f = \frac{\mu_f^{wc} - \mu_f^{oc}}{\sigma_{\text{pooled}}}$$

where μ_f^{wc} and μ_f^{oc} are the mean activations of latent f and σ_{pooled} is the pooled standard deviation from both groups.

Evaluating Associations. Following the work of Simon & Zou (2025), we binarized normalized SAE activations—latent-on/latent-off—using several activation thresholds of 0, 0.1, 0.2, 0.35, 0.5, 0.6, and 0.8. For each pattern, we first evaluated all possible latent-pattern pairs on the validation set and selected those with F1 score ≥ 0.5 to reduce the number of tested patterns. Then, we re-evaluated these pairs on the test set and reported those passing the same criterion. The identified associations can be explored in our interactive visualizer, InterMol.

Metrics. For each latent-pattern-threshold triplet, the association was evaluated using precision, recall, and F1 score, with modified recall and F1 score for substructural patterns. The metrics are defined as follows.

$$\text{Pr} = \frac{TP}{TP + FP} \qquad \text{Re} = \frac{TP}{TP + FN} \qquad \text{F1 score} = \frac{2 \times \text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}$$

$$\text{Re}_{sub} = \frac{TP_{sub}}{n_{sub}} \qquad \text{F1 score}_{sub} = \frac{2 \times \text{Pr} \times \text{Re}_{sub}}{\text{Pr} + \text{Re}_{sub}}$$

where TP_{sub} is the number of substructures where the latent activates on all atoms and n_{sub} is the total number of substructures present in the dataset.

C.2.4. Syntax Parsing Features

Position Latents. In identifying position latents, we used Spearman correlation coefficient, ρ, between the activations and the relative token position within a SMILES string. We set a minimum of 3 activated tokens to calculate the correlation, allowing for rarely-occurring syntactic tokens, such as bond tokens, that still tend to be position latents. To get a latent-wise representative measure, we averaged the Fisher's Z-transformed ρ across the set of molecules in the validation set, transformed them back to Spearman coefficients $\bar{\rho}$, and labeled those with $|\bar{\rho}| \geq 0.6$ as position latents.

Anything-But-X-Latents. Since these latents are expected to be activated on any token, we argue that such latents have dense activation patterns (activation density ≥ 0.2) and distinct non-activated tokens. Activation density was measured by averaging the proportion of activated tokens per SMILES string per latent across the validation set. For distinct non-activated tokens, we used p_{ctx} calculated over the validation set as previously described in Section 4.2.2, with a cutoff of ≥ 0.9 for tokens with total occurrence ≥ 100.

C.3. Chemically Relevant Latents

Datasets. The datasets for each task were fetched using the pytdc package. Most of the 21 tasks used in the study were split using scaffold split, while certain tasks—"Lipophilicity (AstraZeneca)", "Hydration Free Energy (FreeSolv)", "Tox21", and "ClinTox"—used random split. For each task, three different data splits were produced using different random seeds with 80/20 training-test set split.

Probe Features. As baselines, we used Extended-Connectivity Fingerprints (ECFPs; i.e., Morgan Fingerprints) with diameters of 0, 2, and 4, along with 12 physicochemical descriptors: (1) Crippen's octanol-water partition coefficient and (2) molar refractivity, (3) molecular weight, (4) formal charge, (5) fraction of sp^3 hybridized carbon, (6) heavy atom count, (7) total atom count (including implicit hydrogens), (8) ring count, (9) number of hydrogen bond acceptors and (10) donors, (11) number of rotatable bonds, and (12) topological polar surface area (TPSA); all generated using RDKit. The physicochemical descriptors were standardized task-wise over the training set and the mean and variance were reused to standardize the test set features. For MolFormer-10 hidden states and SAE activations across layers, features were obtained via mean- and max-pooling over the SMILES token sequence.

Linear Probes. We trained logistic regression and ridge regression for classification and regression tasks, respectively. All linear probes were optimized using grid search over regularization strengths (10^{-4}, 10^{-2}, 10^{0}, 10^{2}, 10^{4}) with 5-fold cross-validation and evaluated on a separate test set for each data split.

D. Additional Results

D.1. Syntax Parsing Features

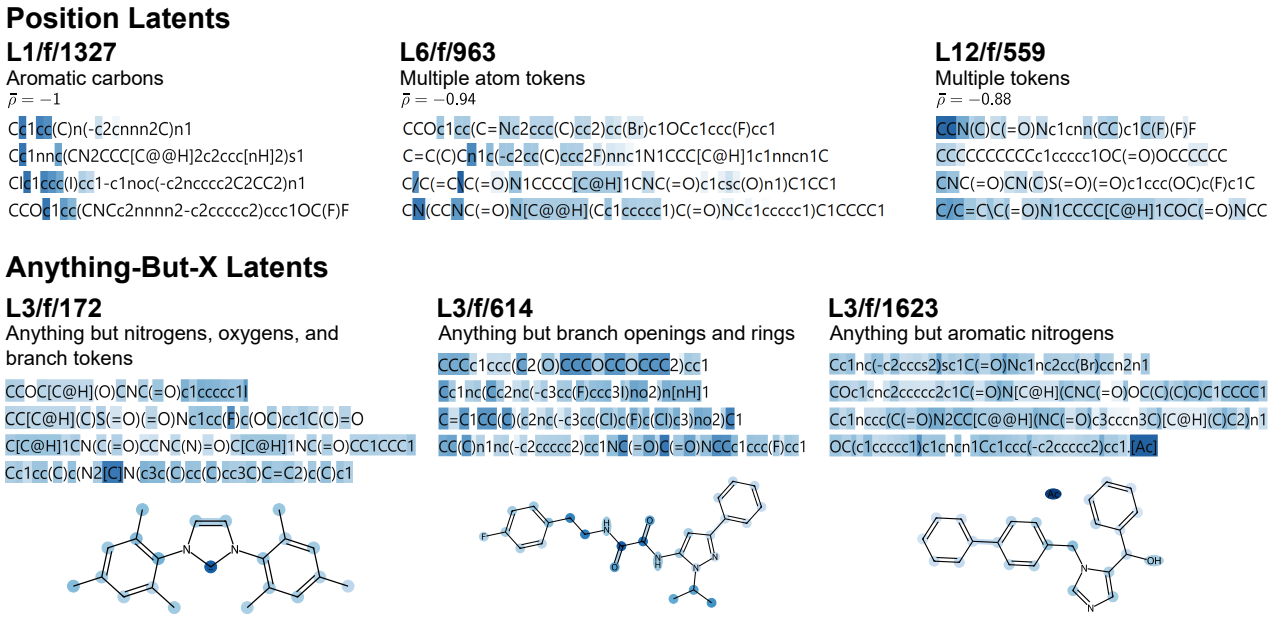

Figure 6. **More Examples of Syntax Parsing Features. Position Latents**: **L1/f/1327** shows decreasing activations on token c in the first few positions; **L3/f/1803** shows decreasing activations on multiple atom tokens; and **L12/f/559** shows decreasing activations across multiple tokens. **Anything-But-X Latents**: **L3/f/172** ($p_{\text{zero}} = 0.97$) does not activate on nitrogens ($p_{\text{ctx}} = 0.99$), oxygens ($p_{\text{ctx}} = 0.99$), branch openings ($p_{\text{ctx}} = 0.99$), and closings ($p_{\text{ctx}} = 0.99$); **L3/f/614** ($p_{\text{zero}} = 0.98$) does not activate on branch openings ($p_{\text{ctx}} = 0.99$) and rings ($p_{\text{ctx}} = 0.99$); and **L3/f/1623** ($p_{\text{zero}} = 0.59$) does not activate on aromatic nitrogens ($p_{\text{ctx}} = 0.99$).

D.2. SMILES Variants Perturb Representations

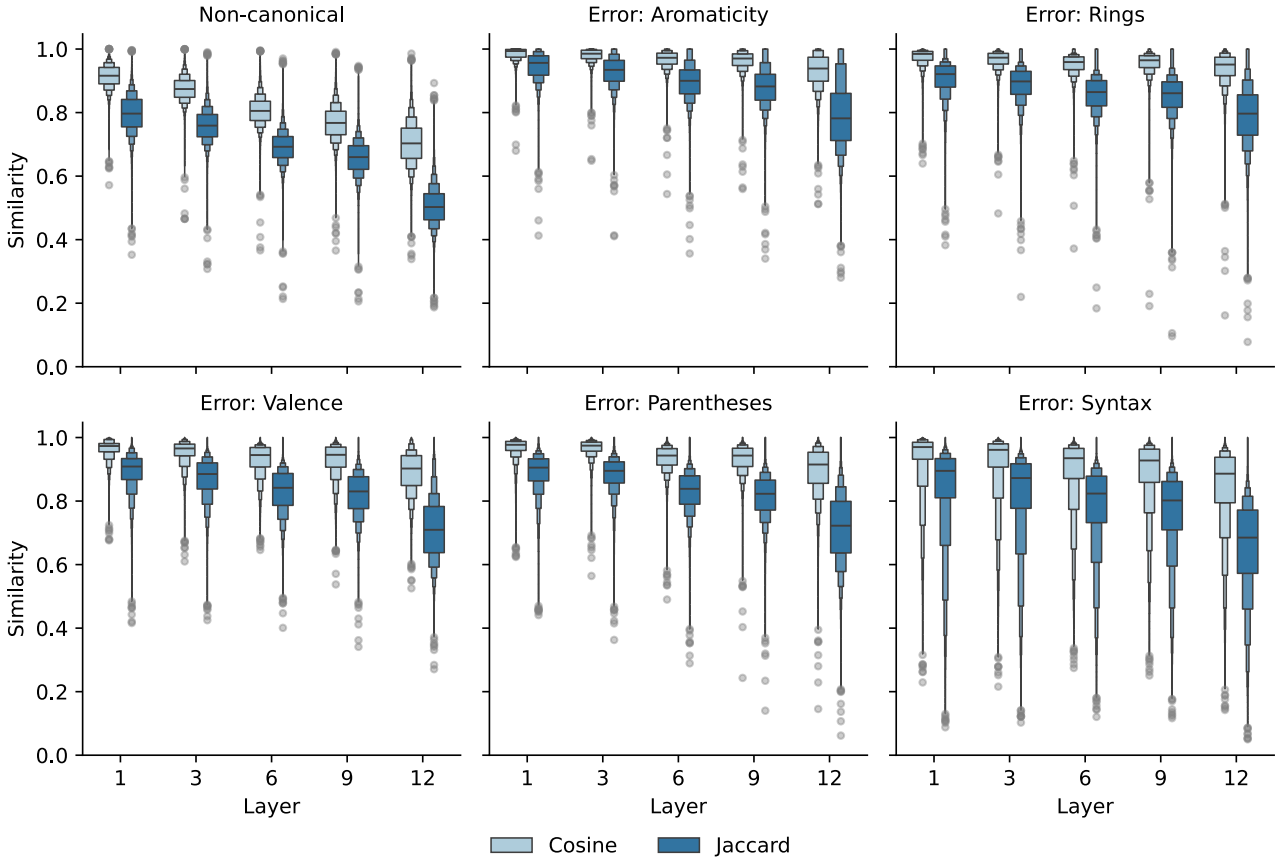

Figure 7. **Distribution of Similarity Measures for Non-canonical and Invalid SMILES.** Generally, we find that the median Jaccard and cosine similarities for invalid SMILES are higher compared to non-canonical SMILES, suggesting that latent representations are more robust to invalid SMILES than to non-canonical SMILES. This is likely because invalid SMILES retain the canonical token order, whereas non-canonical SMILES reorder the atomic traversal and thus shuffle the token sequence. When sorted by median Jaccard and cosine similarity from most to least similar to valid molecules, aromaticity errors are the most similar up to Layer 9, while ring errors being the most similar at the Layer 12. Valence and parentheses errors are typically the third and fourth most similar, with syntax errors being the most perturbed.

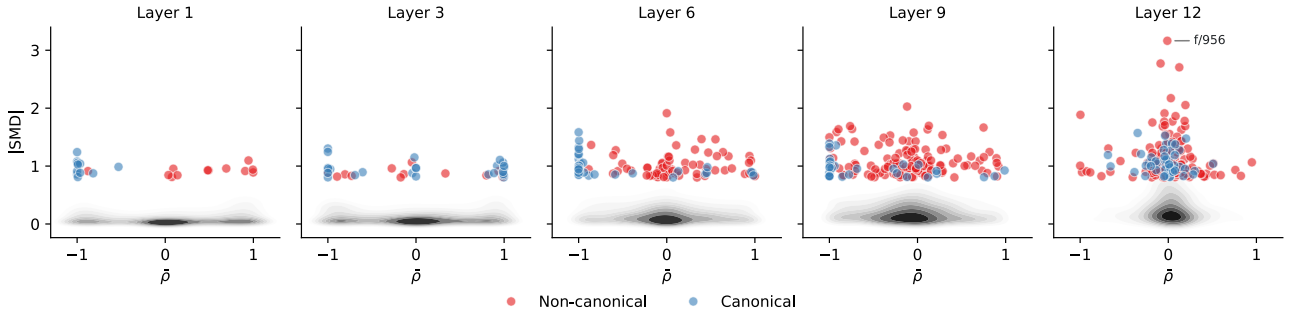

Figure 8. **Non-canonical SMILES: Significant Latents Increase Across Layers.** Grey contour areas show the distribution of latent |SMD| and Spearman correlation to token position $\bar{\rho}$; colored points mark significant latents ($|\text{SMD}| \geq 0.8$) from max-pooled activations, with blue indicating higher |SMD| for canonical and red for non-canonical SMILES. Canonical latents consistently cluster near $\bar{\rho} \approx -1.0$, indicating position latents, but predominate only in early layers. From Layer 6 to Layer 12, the number of significant non-canonical latents exceeds that of canonical ones, likely reflecting how disruption of early-layer position latents alters internal molecular representations in deeper layers.

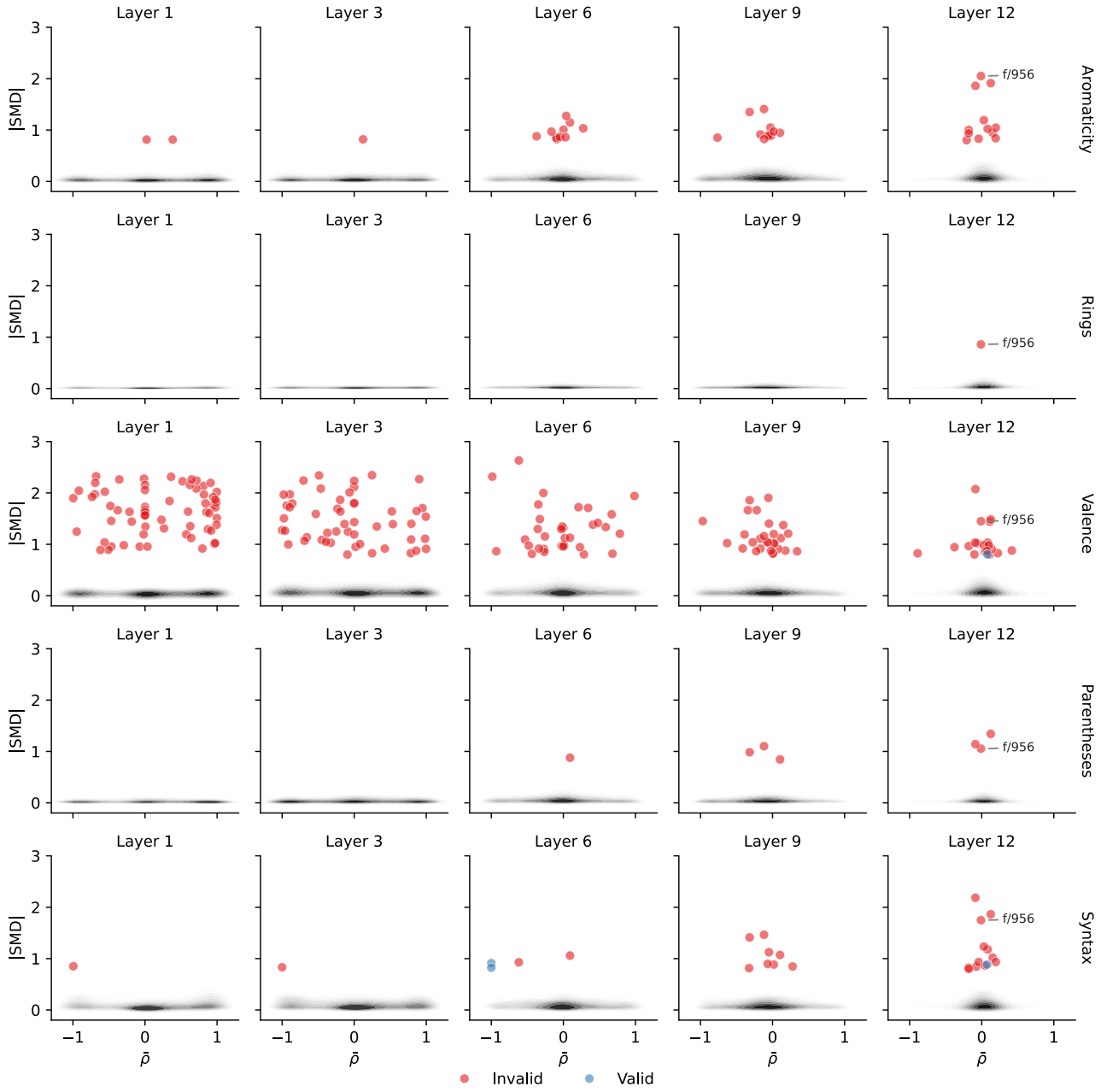

Figure 9. **Invalid SMILES: Valence Errors Have the Most Significant Latents in Early Layers.** Grey contour areas show the distribution of latent |SMD| and Spearman correlation to token position $\bar{\rho}$; colored points mark significant latents ($|\text{SMD}| \geq 0.8$) from max-pooled activations, with blue indicating higher |SMD| for valid and red for invalid SMILES. Unlike augmented SMILES case, high-SMD latents across all error types are mainly driven by stronger activations in invalid SMILES. Interestingly, valence errors have the greatest number of high-SMD latents but are the only error type whose count decreases through layer 12, which may be attributed to how valence errors introduce or modify bond tokens, mostly affecting early-layer latents Table 1.

D.3. Causal Evidence of Position Latent Propagation

We hypothesize that the low final-layer embeddings similarity observed for non-canonical SMILES arises from errors propagating through position latents in early layers. To test this, we ablate these latents by patching their activations with values from the canonical variant, which should restore embedding consistency between non-canonical and canonical SMILES, both representing the same molecule.

Dataset. We use the same dataset as in the augmented SMILES analysis. However, since this analysis requires patching activations from canonical SMILES into their non-canonical counterparts, we filter for non-canonical SMILES with the same token length as their canonical ones. This results in a smaller set of 12,186 canonical SMILES, with a total of 21,545 pairs.

Activation Patching. Activations are patched cumulatively across four configurations: Layer 1; Layer 1 and 3; Layer 1, 3, and 6; and Layer 1, 3, 6, and 9. Unchanged activations are referred to as non-ablated, as shown in Figure 10. In analyzing SAE latent dynamics across layers, we use SMD (see Section 5) for each latent, computed from its max-pooled across tokens, molecule-wise.

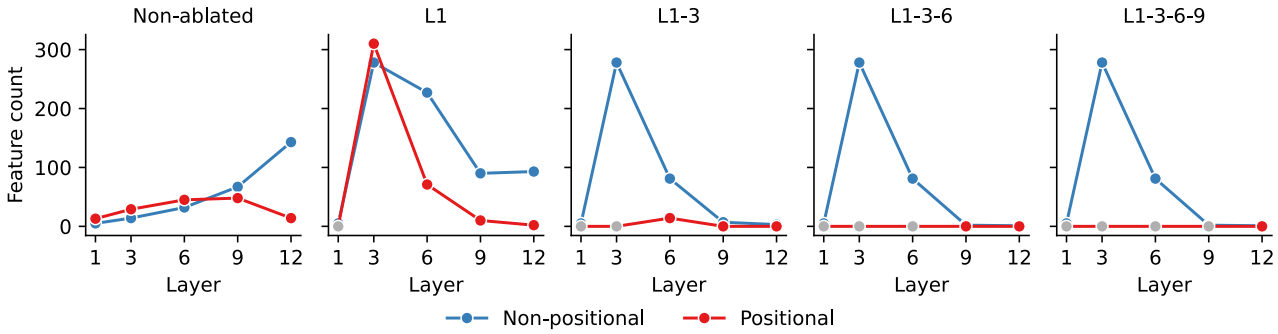

Figure 10. **Early-Layer Position Latents Affect Final-Layer Molecular Representations.** The y-axis represents the number of SAE latents with $|\text{SMD}| \geq 0.8$, and the grey markers indicate the ablated position latents, for which the count is expected to be zero. Activation patching reverses this trend, with layer 3 surging significantly for both positional and non-positional latents upon ablating layer 1. Ablating layers 1 and 3 significantly reduces high-SMD non-positional latents at layer 9 and 12; adding layer 6 drops these further to 1 and 0, respectively. This shows that ablating layers 1 and 3 both reduces downstream high-SMD position latents and corrects later layers' molecular embeddings. Notably, the surge appears at layer 3 and only after ablating layer 1, suggesting layer 3 plays a key role in correcting MolFormer's internal representations.

D.4. Downstream Tasks

Table 2. **Performance on ADMET Classification Tasks.** Values are reported as the mean ± standard deviation of ROC-AUC (↑) across three runs with different random seeds for train-test splits. Tox21 task results are averaged across 12 different targets. Different pooling methods are used to generate MolFormer and SAE molecular features.

Task	Baseline				MolFormer		SAE	
	ECFP0	ECFP2	ECFP4	PcDs[1]	Mean	Max	Mean	Max
Absorption (A)								
Bioavailability (Ma)	0.63 ± 0.07	0.66 ± 0.09	0.68 ± 0.08	0.68 ± 0.08	0.71 ± 0.05 (12)	**0.74 ± 0.07 (9)**	0.72 ± 0.04 (12)	0.71 ± 0.08 (12)
HIA (Hou)	0.95 ± 0.03	0.97 ± 0.01	0.96 ± 0.02	**1.0 ± 0.0**	0.97 ± 0.01 (12)	0.97 ± 0.02 (3)	0.98 ± 0.01 (12)	0.98 ± 0.02 (9)
PAMPA (NCATS)	0.66 ± 0.04	0.71 ± 0.03	0.72 ± 0.01	0.75 ± 0.01	0.75 ± 0.02 (12)	0.75 ± 0.02 (12)	**0.76 ± 0.02 (12)**	0.75 ± 0.03 (12)
Pgp (Broccatelli)	0.77 ± 0.06	0.91 ± 0.03	0.91 ± 0.02	0.91 ± 0.0	0.92 ± 0.01 (6)	0.92 ± 0.02 (12)	0.92 ± 0.0 (1)	**0.93 ± 0.01 (12)**
Distribution (D)								
BBB (Martins)	0.81 ± 0.04	0.88 ± 0.03	0.86 ± 0.05	0.81 ± 0.01	0.87 ± 0.02 (6)	0.86 ± 0.01 (12)	0.88 ± 0.02 (6)	**0.9 ± 0.03 (6)**
Metabolism (M)								
CYP2C9 (Veith)	0.73 ± 0.03	0.86 ± 0.01	0.86 ± 0.01	0.8 ± 0.01	0.86 ± 0.01 (12)	0.87 ± 0.01 (12)	0.87 ± 0.01 (12)	**0.88 ± 0.01 (12)**
CYP2D6 (Veith)	0.75 ± 0.02	0.85 ± 0.0	0.84 ± 0.01	0.77 ± 0.01	0.85 ± 0.0 (6)	0.84 ± 0.01 (12)	**0.86 ± 0.0 (6)**	0.85 ± 0.01 (12)
Toxicity (T)								
AMES	0.7 ± 0.02	0.81 ± 0.01	0.79 ± 0.01	0.73 ± 0.01	0.8 ± 0.01 (12)	0.81 ± 0.0 (12)	0.82 ± 0.01 (12)	**0.82 ± 0.02 (9)**
Carcinogens (Lagunin)	0.73 ± 0.17	0.72 ± 0.22	0.72 ± 0.21	0.73 ± 0.16	0.74 ± 0.07 (1)	0.75 ± 0.11 (3)	0.74 ± 0.19 (1)	**0.75 ± 0.13 (1)**
ClinTox	0.9 ± 0.04	0.87 ± 0.05	0.82 ± 0.07	0.77 ± 0.05	0.9 ± 0.03 (9)	0.91 ± 0.01 (12)	0.9 ± 0.03 (6)	**0.91 ± 0.02 (6)**
Tox21[2]	0.74 ± 0.04	0.81 ± 0.05	0.79 ± 0.05	0.76 ± 0.05	0.81 ± 0.05 (9)	0.81 ± 0.04 (9)	**0.82 ± 0.05 (9)**	0.81 ± 0.05 (12)

[1]Physicochemical descriptors.
[2]Reported scores are also averaged across multiple binary targets.

Table 3. **Performance on ADMET Regression Tasks.** Values are reported as the mean ± standard deviation of RMSE (↓) across three runs with different random seeds for train-test splits. Different pooling methods are used to generate MolFormer and SAE molecular features.

Task	Baseline				MolFormer		SAE	
	ECFP0	ECFP2	ECFP4	PcDs[1]	Mean	Max	Mean	Max
Absorption (A)								
Caco2 (Wang)	0.66 ± 0.06	0.59 ± 0.03	0.56 ± 0.02	0.54 ± 0.01	**0.49 ± 0.02 (12)**	0.52 ± 0.06 (9)	0.51 ± 0.03 (12)	0.5 ± 0.0 (12)
Hydration Free Energy (FreeSolv)	2.26 ± 0.36	1.39 ± 0.34	1.6 ± 0.37	1.5 ± 0.05	1.38 ± 0.08 (3)	1.51 ± 0.12 (1)	1.53 ± 0.23 (6)	**1.2 ± 0.19 (1)**
Lipophilicity (AstraZeneca)	1.11 ± 0.07	0.81 ± 0.03	0.85 ± 0.05	1.03 ± 0.06	0.78 ± 0.04 (12)	0.81 ± 0.03 (12)	**0.75 ± 0.04 (12)**	0.75 ± 0.04 (12)
Solubility (AqSolDB)	2.39 ± 0.07	2.02 ± 0.21	1.87 ± 0.13	2.21 ± 0.27	1.88 ± 0.59 (3)	1.77 ± 0.41 (1)	1.83 ± 0.57 (6)	**1.65 ± 0.35 (1)**
Distribution (D)								
PPBR (AZ)	14.95 ± 1.61	13.4 ± 1.46	13.25 ± 1.38	13.72 ± 1.28	12.79 ± 0.76 (12)	13.3 ± 1.15 (6)	12.75 ± 1.05 (9)	**12.66 ± 0.99 (6)**
VDss (Lombardo)	20.7 ± 23.56	20.55 ± 23.41	20.54 ± 23.44	20.44 ± 23.66	20.44 ± 23.75 (12)	**20.26 ± 23.49 (6)**	20.34 ± 23.54 (6)	20.32 ± 23.64 (1)
Excretion (E)								
Clearance Hepatocyte (AZ)	49.11 ± 1.87	46.78 ± 1.3	46.57 ± 0.54	49.8 ± 1.76	46.47 ± 0.5 (12)	47.11 ± 1.2 (1)	46.51 ± 1.81 (9)	**45.58 ± 1.25 (12)**
Clearance Microsome (AZ)	43.13 ± 1.63	39.08 ± 1.88	39.6 ± 1.86	44.14 ± 1.42	39.02 ± 2.12 (12)	39.89 ± 1.05 (9)	38.49 ± 1.63 (12)	**37.43 ± 2.06 (12)**
Half Life (Obach)	23.74 ± 5.2	29.38 ± 5.87	30.1 ± 5.08	38.57 ± 30.37	19.81 ± 2.4 (12)	**19.66 ± 2.91 (6)**	26.0 ± 8.99 (9)	25.64 ± 8.21 (3)
Toxicity (T)								
LD50 (Zhu)	0.85 ± 0.03	0.81 ± 0.06	0.79 ± 0.02	0.91 ± 0.02	0.81 ± 0.02 (6)	0.83 ± 0.03 (1)	0.8 ± 0.01 (6)	**0.78 ± 0.02 (6)**

[1]Physicochemical descriptors.

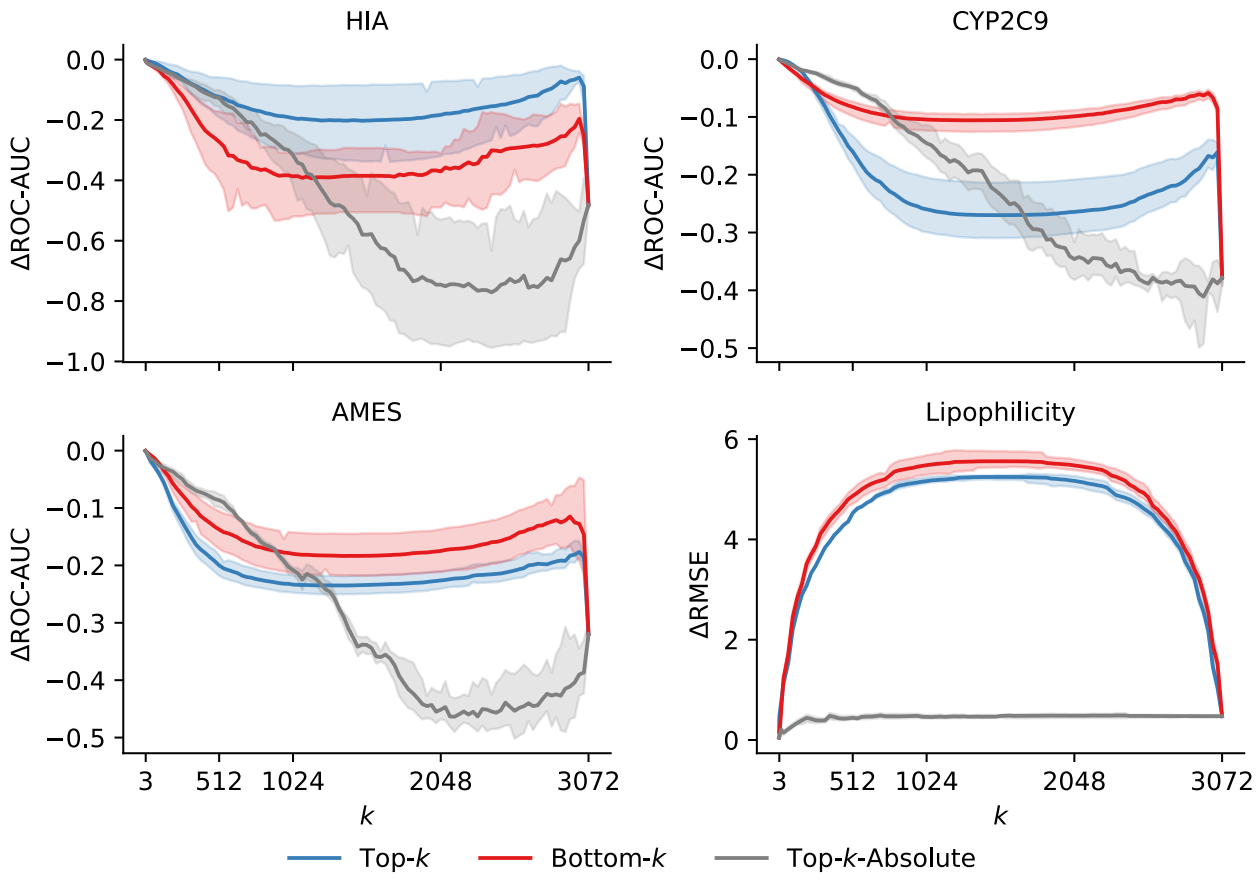

Figure 11. **Pharmacological Encoding Is Shared Across Multiple SAE Latents.** Shown here is the change in linear probe performance upon ablating latents ranked by $z(\beta_f)$ – top-k, bottom-k, and top-k by absolute value – on the test set of each data split. Ablating the top-3, bottom-3, or even top-3-absolute latents does not significantly degrade performance, whereas ablating 512 or more latents does, indicating that task encoding is distributed across many SAE latents. Across both top-k and bottom-k ablations, classification and regression tasks show similar performance degradation trends, except that classification tasks exhibit a slight recovery at very large k. Meanwhile, the regression task produces symmetric ΔRMSE curves, since ΔRMSE is evaluated in absolute value; the top-k-absolute curve still suggests similar behavior to classification probes.

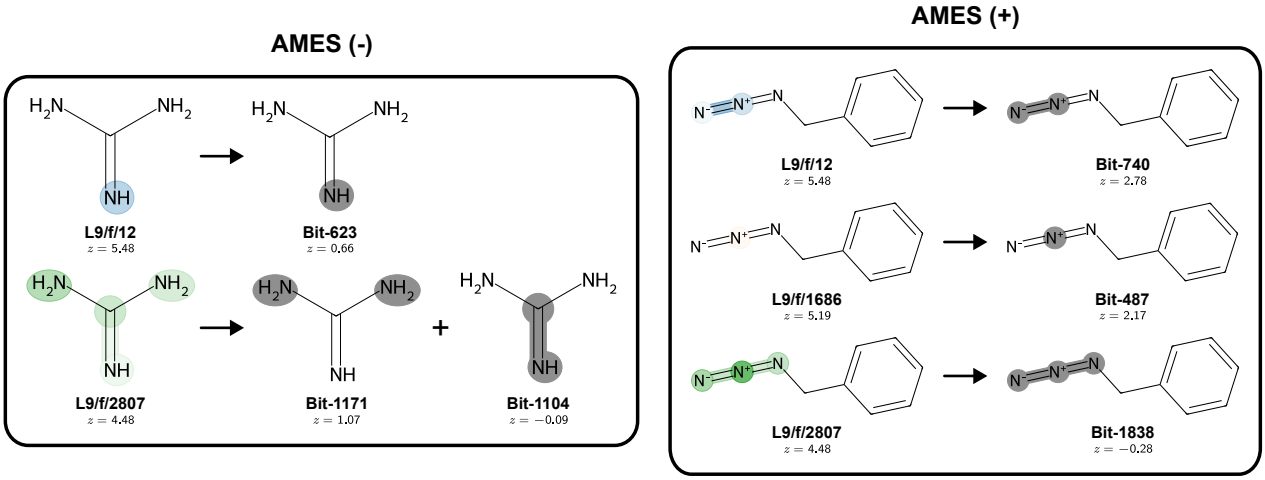

Figure 12. **SAE Latents Align with Morgan Fingerprint Features.** Two molecules from the AMES dataset show that concepts encoded in the SAE latents can correspond to ECFP2 bits. A single SAE latent may be correspond to a fraction of, exactly one, or many ECFP2 bits. Moreover, SAE latents offer not only binary representations but also activation magnitudes, as shown by L9/f/2807, which activates strongly at the `[N+]` atom in the AMES (+) molecule. This richer representation may also explain the slightly higher evaluation performance of SAE latents, particularly when using max-pooling features, over ECFP on most downstream tasks.