Unveiling Latent Knowledge in Chemistry Language Models through Sparse Autoencoders

Jaron Cohen[1,*] **Alexander G. Hasson**[2,*] **Sara Tanovic**[3,*]

[1]Independent researcher
[2]Department of Oncology, University of Oxford [3]Department of Chemistry, University of Oxford
jdicohen@gmail.com alexander.hasson@gtc.ox.ac.uk sara.tanovic@gtc.ox.ac.uk

Abstract

Since the advent of machine learning, interpretability has remained a persistent challenge, becoming increasingly urgent as generative models support high-stakes applications in drug and material discovery. Recent advances in large language model (LLM) architectures have yielded chemistry language models (CLMs) with impressive capabilities in molecular property prediction and molecular generation. However, how these models internally represent chemical knowledge remains poorly understood. In this work, we extend sparse autoencoder techniques to uncover and examine interpretable features within CLMs. Applying our methodology to the Foundation Models for Materials (FM4M) SMI-TED chemistry foundation model, we extract semantically meaningful latent features and analyse their activation patterns across diverse molecular datasets. Our findings reveal that these models encode a rich landscape of chemical concepts. We identify correlations between specific latent features and distinct domains of chemical knowledge, including structural motifs, physicochemical properties, and pharmacological drug classes. Our approach provides a generalisable framework for uncovering latent knowledge in chemistry-focused AI systems. This work has implications for both foundational understanding and practical deployment; with the potential to accelerate computational chemistry research.

1 Introduction

The intersection of artificial intelligence (AI) and chemistry has recently witnessed unprecedented advances with the emergence of foundational chemistry language models (CLMs)[1,2]. Built upon Transformer[3] architectures, these models have been fine-tuned for tasks in molecular property prediction and *de novo* materials design, often matching or exceeding traditional approaches.[4,5,6,7] Yet, these empirical successes come with a critical limitation: the models operate as "black boxes," their internal decision-making processes opaque to human understanding. This interpretability challenge is particularly acute as it touches on a fundamental epistemological question: are these models learning genuine chemical principles, or are they sophisticated pattern-matching systems?

Without interpretable representations, we cannot distinguish between models that have internalised the physical laws governing molecular behaviour and those that merely memorise statistical correlations in training data. This distinction has profound implications for model generalisation, scientific discovery, and the regulatory approval of AI-assisted therapeutics. The core difficulty lies in deciphering the holistic, molecular-level vectors that represent entire chemical structures, where concepts are entangled and distributed.

Recent advances in sparse autoencoders (SAEs) [8,9,10,11,12] provide a promising path toward interpretability. SAEs decompose neural network activations into sparse *features* that can correspond

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Design.

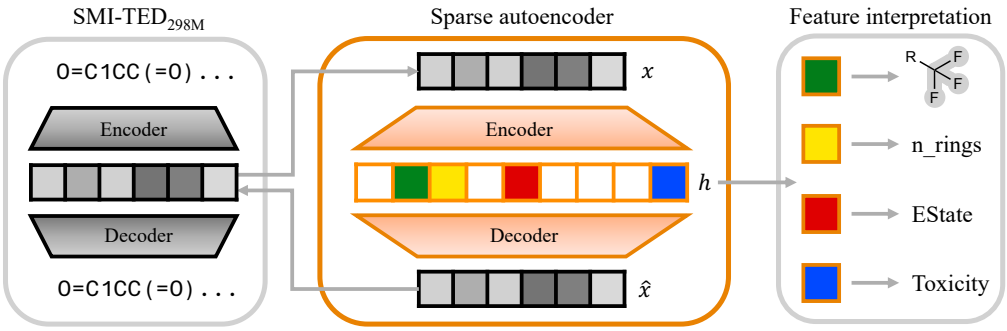


Figure 1: Overview of our workflow. Embeddings are extracted from SMI-TED and converted into features via the SAE model. These features are then interpreted to find relationships with structural and physical information.

to interpretable and meaningful concepts.[9] However, their application to chemistry has remained unexplored. **In this work, we present the first application of SAEs to CLMs**, specifically the state-of-the-art SMI-TED foundation model[7], presenting the first systematic investigation of interpretable features within CLMs. We train SAEs on the internal representations of the model and analyse the resulting sparse features across diverse molecular datasets (Figure 1). Our analysis reveals that these models develop rich, hierarchical representations of chemical knowledge, with individual features corresponding to structural motifs, physicochemical properties, and pharmacological drug classes - concepts never explicitly provided during self-supervised training.

Contributions **(1)** The first application of SAE techniques to CLMs, revealing interpretable chemical features within foundation model representations. **(2)** A novel domain-specific evaluation framework that validates chemical interpretability through molecular descriptors, substructure analysis, and functional annotations. **(3)** Demonstration of feature steering capabilities that enable causal manipulation of molecular representations while preserving chemical validity. All materials needed to reproduce our results including model weights will be made available at the time of publication.

2 Methodology

2.1 Problem Formulation and Model Setup

We formalise the problem of interpreting CLMs via sparse dictionary learning.[8] Our central hypothesis is that a dense molecular representation vector, $\mathbf{x} \in \mathbb{R}^{d_{\text{model}}}$, can be sparsely decomposed as $\mathbf{x} \approx \sum_i h_i \mathbf{w}_i$, where $\{\mathbf{w}_i\}$ is a dictionary of interpretable feature vectors and $\mathbf{h}$ is a sparse activation vector. We extract these fixed-size vectors ($d_{\text{model}} = 768$) from the *submersion layer* of the SMI-TED foundation model, which is responsible for mapping a sequence of molecular tokens into a single, fixed-size vector that represents the entire molecule.

2.2 SAE Architecture Training and Evaluation

To learn this decomposition, we implement a TopK SAE[13]. We select this architecture over traditional L_1-regularised approaches due to its direct control over feature sparsity and its demonstrated ability to achieve a superior fidelity-sparsity trade-off[13]. The encoder identifies the k most active features for a given input, and the decoder then attempts to reconstruct the original vector using only this small subset.

We train our SAEs on a dataset of 5 million molecular representations extracted from SMI-TED. We curate this data from PubChem following the filtering and preparation protocol described by Soares *et al.*[14] to ensure a highly similar data distribution (see Appendix S2.1.1). The training objective is to minimise reconstruction loss ($\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$) while balancing sparsity and feature utilisation.

To identify an optimal model configuration, we perform a grid search over key hyperparameters, including the dictionary expansion factor (the size of the feature dictionary relative to the input dimension, e.g., 8×, 16×, 32×) and the sparsity level $k \in \{40, 80, 160\}$. This process allows us to

map the Pareto frontier of models that optimally trade off reconstruction fidelity for sparsity. From this frontier, we select a final model that demonstrates the most promising initial signs of feature interpretability while ensuring its reconstructed vectors can still be successfully decoded back into their original, chemically valid SMILES string (see Appendix S2.3 for further details).

We additionally prepare the ChEMBL35 [15], MITOTOX [16], and MOSES [17] datasets as per our filtering and preparation protocol to investigate the interpretability of physicochemical, functional, and structural concepts (see Appendix S2.1.2).

3 Results

We begin our analysis by profiling the features from our selected SAE ($8\times$ expansion, $k = 80$), which is used for all subsequent investigations. We construct the *feature landscape* in Figure 2, which maps each feature's activation frequency, intensity, and volatility. This visualisation reveals a spectrum of feature types from specialist features (left, rare but specific) to generalist features (right, common but potentially polysemantic). For example, features 247 and 266 selectively detect specific chemical substructures (highlighted in white), activating rarely but consistently across molecules sharing these motifs, while features 80 and 429 activate frequently across structurally diverse molecules, suggesting they encode broader chemical concepts or multiple properties.

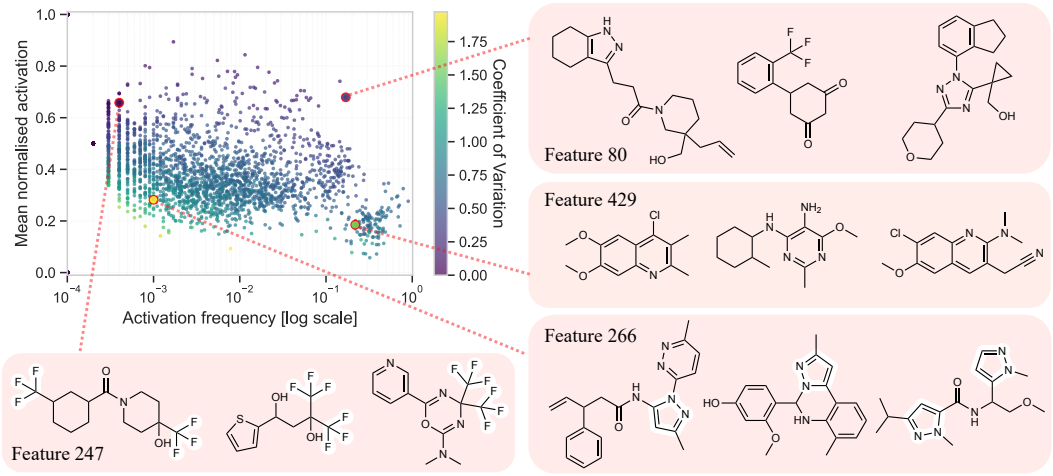


Figure 2: Feature landscape calculated on a 10k subset of the MOSES dataset. Each point represents one SAE feature, mapped according to three metrics: (1) Activation frequency (x-axis) measures how many molecules activate this feature, revealing whether it detects common or rare chemical attributes; (2) Mean normalised activation (y-axis) quantifies the typical strength when the feature activates, indicating its importance when present; (3) Coefficient of variation (colour gradient) represents consistency of activation strength, with darker points showing more consistent behaviour.

3.1 Substructures

We test the claim that SAEs produce an interpretable feature vocabulary by evaluating if individual features detect chemical concepts more effectively than individual neurons. A comparison of max F1 scores for 14 functional groups (Table S3) shows that SAE features outperform neurons. This result indicates that the features form a disentangled representation, isolating specific informational components previously distributed across the latent space.

The performance gap between features and neurons is largest for motifs with low prevalence in the training data. For instance, nitrate groups appear in only 5,167 of the 5 million molecules (0.1%). The feature for detecting this rare group achieves a perfect F1 score of 1.000, while the best neuron scores only 0.056. Large differences also exist for other low-prevalence motifs like acetylenic carbon (0.933 vs. 0.079; 0.8% prevalence) and cyanamide (0.667 vs. 0.030; 0.1% prevalence). These results suggest the SAE constructs new detectors from linear combinations of neurons, a necessary step when the base model avoids dedicating single neurons to rare concepts. Visualisations of top-activating

molecules for these features confirm their precision, as they activate exclusively on molecules with the target substructure (Figure S4).

To establish a causal link between a feature and its correlated motif, we perform feature steering via ablation. For a given molecule, we set the activation of a target feature to zero before decoding the molecule from its modified latent representation. Examples of this ablation experiment for three features are provided in Figure 3. This intervention produces a targeted and predictable modification; for example, ablating the feature for a carbonyl group (Feature 758) selectively removes that group from the molecular structure while preserving the core scaffold, and instead replaces it with a pentyl chain. This result provides direct evidence that the feature causally encodes the information required to generate the substructure, confirming its functional role in the model's generative process.

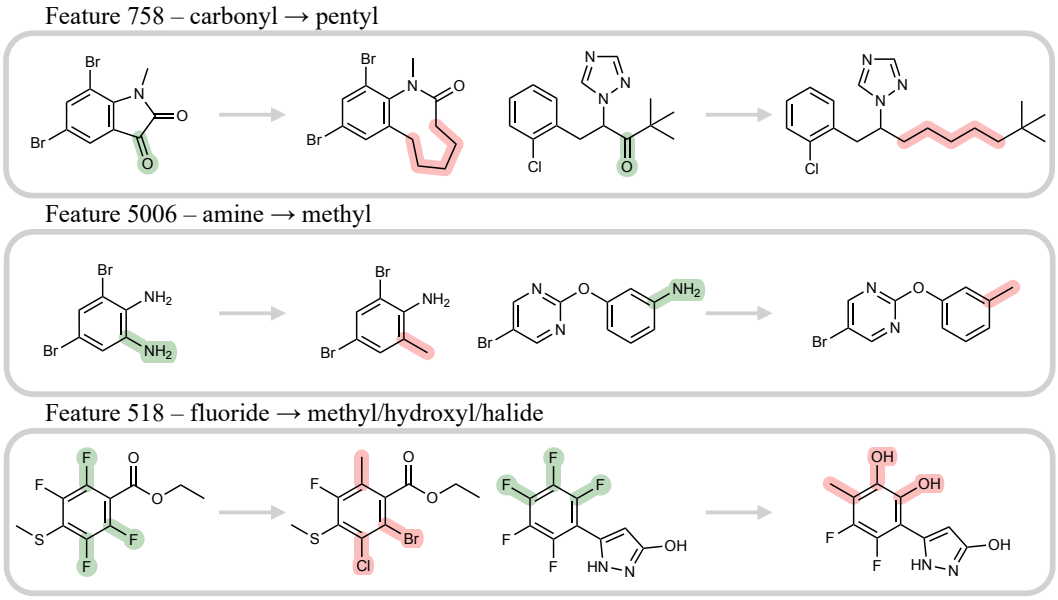


Figure 3: Examples of molecules with altered substructures highlighted before (green) and after (red) steering. Steering is performed by setting the specified feature activation to zero.

3.2 Physicochemical Properties

Beyond local substructures, SAE features also provide a more disentangled representation of global physicochemical properties. Comparing against raw neurons, Principal Component Analysis (PCA) components, and Non-negative Matrix Factorisation (NMF) factors, we find each descriptor correlates (Spearman's $\rho \geq 0.3$) with 6.40 ± 3.78 features, compared to 65.77 ± 51.07 neurons, 1.08 ± 1.29 PCA components, and 3.70 ± 5.37 NMF factors. While PCA appears most parsimonious, the 100 strongest descriptor relationships map to only 3 different PCA components versus 100 unique SAE features, 5 NMF factors, and 69 unique neurons. This suggests PCA efficiently captures variance but conflates multiple chemical concepts within single dimensions. The SAE representation's lower redundancy is confirmed by the mean pairwise correlation among features (0.016 ± 0.029), which is an order of magnitude smaller than that of neurons (0.162 ± 0.122). The top three correlated features are visualised in Figure S5, wherein steering of the top three activated molecules shows the causal relationship between activation and descriptor. These results demonstrate that the SAE distils the original neuron activations into a set of compact and decorrelated features, providing a more interpretable basis for aligning the model's internal representations with external physicochemical descriptors than either the original neurons or standard dimensionality reduction techniques.

3.3 Functional Behaviour

Sections 3.1 and 3.2 establish that SAE features form a disentangled basis for substructures and physicochemical properties. We now investigate if these features also represent higher-level functional concepts. We first use a downstream prediction task, toxicity prediction on the MITOTOX dataset, to compare the utility and efficiency of the SAE feature basis against the original neuron basis.

We train multiple logistic regression models on both representations and find their predictive performance to be nearly identical. Models trained on the 768 neuron activations yield a mean AUCpr of 0.603 ± 0.035, while models trained on the 6144 SAE features yield 0.606 ± 0.034, a statistically insignificant difference ($t = -0.34$, $p = 0.75$). This result confirms that the SAE decomposition is information-preserving for this task. However, the models differ in their sparsity. The logistic regression procedure identifies 213 neurons (0.277% of variables) as significant predictors, whilst requiring only 19 SAE features (0.003% of variables) to achieve the same performance (the top 3 of which are visualised in Figure S7). This finding suggests the SAE isolates the relevant biological signal into a more compact set of features.

We now present a case study that moves from this multi-feature abstraction to a single feature associated with a specific pharmacological mechanism. We identify a feature that activates selectively for three compounds – CHEMBL1672485[18], CHEMBL454618[19], and CHEMBL368522[20] – which span distinct chemotypes (Figure S8). While two of the molecules are very close structural analogues, the third exhibits relatively weak structural similarity to them, as measured by a Tanimoto similarity that falls within the range expected under a null distribution of random molecular pairs (see Appendix S4.3). There are molecules in the dataset that have higher Tanimoto similarities between all three molecules that this feature does not activate for.

All three compounds have been reported to share activity for the μ-, κ-, and δ-opioid receptors, according to both ZINC SEA[21] *in silico* predictions and experimental ChEMBL bioactivity data. The maximum common substructure (MCS) across the three molecules is also present in 247 other molecules, suggesting that this feature cannot be explained by the MCS alone. This contrasts with the majority of features, which appear to correspond to small, local structural motifs.

The presence of a feature that activates across both a morphinan-like scaffold (CHEMBL1672485) and a distinct polycyclic scaffold family (CHEMBL454618/368522) suggests that the model is not solely encoding local topological similarity, but instead captures higher-order abstractions that relate to, though do not perfectly determine, shared pharmacological behaviour. The emergence of such a latent feature points toward the model's internal representations being sensitive to patterns that integrate both structural and functional information across chemotypes, rather than reflecting purely structural or purely functional groupings.

4 Conclusion

In this work, we demonstrate that SAEs can decompose the latent representations of a CLM into a more interpretable feature basis. This disentanglement reveals a rich landscape of features spanning local substructural motifs, global physicochemical descriptors, and high-level functional concepts. For instance, we show that the SAE isolates the signal for toxicity into a far more compact representation than the neuron basis, and that single features can group molecules by a shared pharmacological function. Crucially, we show these features are also causally relevant, enabling targeted, step-wise modifications to molecular structures through simple interventions – a capability not afforded by the original, entangled neuron basis. This work provides evidence that the model's internal representations encode a rich hierarchy of functionally relevant chemical concepts and offers a path toward more controllable and interpretable models.

Our approach, however, has several limitations. The interpretation of features remains a presently unscaled process, and our analysis is confined to a single model architecture. The features discovered are also contingent on the training data's chemical space, and their generalisation to out-of-distribution molecules remains an open question. Furthermore, while we compare our features to the raw neuron basis, more robust baselines are needed to fully validate the decomposition's effectiveness. Future work should focus on four key directions. First, developing methods for the automated interpretation of chemical features. Second, exploring applications in AI safety, where identifying and ablating features correlated with undesirable properties (e.g., toxicity) could make models safer. Third, investigating how features behave across different model architectures and scales. This includes exploring concepts like *feature splitting*[22], where a single feature at one SAE scale may decompose into more fundamental sub-features at a finer scale. Finally, moving beyond simple ablation to more sophisticated generative control, such as feature arithmetic, to enable multi-objective molecular optimisation[23]. This work provides a foundational step toward building more transparent, trustworthy, and controllable models for accelerated materials discovery.

References

[1] M. Moret, I. Pachon Angona, L. Cotos, S. Yan, K. Atz, C. Brunner, M. Baumgartner, F. Grisoni and G. Schneider, *Nature Communications*, 2023, **14**, 114.

[2] R. Özçelik, S. De Ruiter, E. Criscuolo and F. Grisoni, *Nature Communications*, 2024, **15**, 6176.

[3] A. Vaswani, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser and I. Polosukhin, 31st Conference on Neural Information Processing Systems, 2017.

[4] R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 15022.

[5] W. Ahmad, E. Simon, S. Chithrananda, G. Grand and B. Ramsundar, *ChemBERTa-2: towards chemical foundation models*, 2022, `http://arxiv.org/abs/2209.01712`, arXiv:2209.01712 [cs].

[6] C. Edwards, T. Lai, K. Ros, G. Honke, K. Cho and H. Ji, *Translation between molecules and natural language*, 2022, `http://arxiv.org/abs/2204.11817`, arXiv:2204.11817 [cs].

[7] E. Soares, E. Vital Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira and K. Schmidt, *Commun. Chem.*, 2025, **8**, 193.

[8] B. A. Olshausen and D. J. Field, *Vision Research*, 1997, **37**, 3311–3325.

[9] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan and C. Olah, *Transformer Circuits Thread*, 2023.

[10] H. Cunningham, A. Ewart, L. Riggs, R. Huben and L. Sharkey, *Sparse Autoencoders Find Highly Interpretable Features in Language Models*, 2023.

[11] S. Rajamanoharan, A. Conmy, L. Smith, T. Lieberum, V. Varma, J. Kramár, R. Shah and N. Nanda, *Improving Dictionary Learning with Gated Sparse Autoencoders*, 2024.

[12] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah and T. Henighan, *Transformer Circuits Thread*, 2024.

[13] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike and J. Wu, *ArXiv*, 2024, **abs/2406.04093**, year.

[14] E. Soares, E. V. Brazil, V. Shirasuna, D. Zubarev, R. Cerqueira and K. Schmidt, *Communications Chemistry*, 2025, **8**, 193.

[15] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. de Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, M. P. Magarinos, N. Bosc, R. Arcila, T. Kizilören, A. Gaulton, A. P. Bento, M. F. Adasme, P. Monecke, G. A. Landrum and A. R. Leach, *Nucleic Acids Research*, 2023, **52**, D1180–D1192.

[16] Y.-T. Lin, K.-H. Lin, C.-J. Huang and A.-C. Wei, *BMC Bioinformatics*, 2021, **22**, 369.

[17] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik and A. Zhavoronkov, *Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models*, 2020.

[18] T. Nemoto, N. Yamamoto, A. Watanabe, H. Fujii, K. Hasebe, M. Nakajima, H. Mochizuki and H. Nagase, *Bioorganic Medicinal Chemistry*, 2011, **19**, 1205–1221.

[19] S. Sakami, K. Kawai, M. Maeda, T. Aoki, H. Fujii, H. Ohno, T. Ito, A. Saitoh, K. Nakao, N. Izumimoto, H. Matsuura, T. Endo, S. Ueno, K. Natsume and H. Nagase, *Bioorganic Medicinal Chemistry*, 2008, **16**, 7956–7967.

[20] J. Clayson, A. Jales, R. J. Tyacke, A. L. Hudson, D. J. Nutt, J. W. Lewis and S. M. Husbands, *Bioorganic Medicinal Chemistry Letters*, 2001, **11**, 939–943.

[21] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.

[22] D. Chanin, J. Wilken-Smith, T. Dulka, H. Bhatnagar, S. Golechha and J. Bloom, *A Is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders*, 2025.

[23] M. Ansari, J. Watchorn, C. E. Brown and J. S. Brown, *dZiner: Rational Inverse Design of Materials with AI Agents*, 2024.

[24] L. Bereska and E. Gavves, *Mechanistic Interpretability for AI Safety – A Review*, 2024.

[25] T. Räuker, A. Ho, S. Casper and D. Hadfield-Menell, *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*, 2023.

[26] E. Simon and J. Zou, *InterPLM: discovering interpretable features in protein language models via sparse autoencoders*, 2025, https://www.biorxiv.org/content/10.1101/2024.11.14.623630v2, Pages: 2024.11.14.623630 Section: New Results.

[27] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.

[28] S. Poux, C. N. Arighi, M. Magrane, A. Bateman, C.-H. Wei, Z. Lu, E. Boutet, H. Bye-A-Jee, M. L. Famiglietti, B. Roechert and T. UniProt Consortium, *Bioinformatics*, 2017, **33**, 3454–3460.

[29] N. Parsan, D. J. Yang and J. J. Yang, *Towards interpretable protein structure prediction with sparse autoencoders*, 2025, http://arxiv.org/abs/2503.08764, arXiv:2503.08764 [q-bio].

[30] J. Su, Y. Lu, S. Pan, A. Murtadha, B. Wen and Y. Liu, *RoFormer: Enhanced Transformer with Rotary Position Embedding*, 2023, https://arxiv.org/abs/2104.09864.

[31] B. Bussmann, P. Leask and N. Nanda, *BatchTopK Sparse Autoencoders*, 2024.

[32] B. Bussmann, N. Nabeshima, A. Karvonen and N. Nanda, *Learning Multi-Level Features with Matryoshka Sparse Autoencoders*, 2025, https://arxiv.org/abs/2503.17547.

[33] *GitHub - JacksonBurns/chemeleon_tox — github.com*, https://github.com/JacksonBurns/chemeleon_tox/tree/main, 2025.

[34] *GitHub - JacksonBurns/mordred-community: Community-Maintained Version of mordred — github.com*, https://github.com/JacksonBurns/mordred-community.

[35] H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *Journal of Cheminformatics*, 2018, **10**, 4.

[36] A. K. Samuel Marks and A. Mueller, *dictionary_learning*, https://github.com/saprmarks/dictionary_learning, 2024.

[37] IBM Research, *materials.smi-ted Model*, https://huggingface.co/ibm-research/materials.smi-ted, 2024.

[38] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramár, A. Dragan, R. Shah and N. Nanda, *Gemma scope: open sparse autoencoders everywhere all At once on gemma 2*, 2024, http://arxiv.org/abs/2408.05147, arXiv:2408.05147 [cs].

[39] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, 2014.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

[41] *Daylight>SMARTS Examples — daylight.com*, https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html.

[42] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.

S1 Background and Related Work

S1.1 Mechanistic Interpretability

Mechanistic interpretability is an approach to reverse engineering neural networks with the goal of understanding their internal computational mechanisms [24,25]. A central challenge in this field is the *superposition hypothesis*, which posits that models learn to represent more features than they have neurons, compressing information efficiently. This compression is thought to give rise to *polysemanticity*, where a single neuron activates for multiple, seemingly unrelated concepts, thus obscuring its specific functional role. Sparse Autoencoders (SAEs) have emerged as a promising methodology for addressing this issue[8,9,10,11,12]. By training an autoencoder to reconstruct a model's internal activations from a sparse, overcomplete feature dictionary, SAEs attempt to disentangle these superimposed representations. The intended outcome is the identification of *putatively monosemantic* features, where each feature vector ideally corresponds to a single, human-interpretable concept, thereby rendering the model's learned knowledge more amenable to systematic analysis.

S1.2 Chemistry Foundation Models

Foundational chemistry language models (CLMs) have evolved from natural language processing and treat Simplified Molecular Input Line Entry System (SMILES) strings as the language of chemistry. State-of-the-art foundation models such as Chemformer,[4] ChemBERTa,[5] MolT5,[6], and SMI-TED[7] are pretrained via self-supervised learning on large databases of SMILES strings for reconstruction tasks. Models can then be fine-tuned with smaller labelled datasets towards specific chemical tasks, such as property prediction, *de novo* molecular design, or retrosynthesis prediction.

S1.3 Related Work in Biological Sequence Models

The application of SAEs to biological language models provides the closest precedent to our work. Simon and Zou[26] trained SAEs on ESM-2[27] embeddings, successfully extracting interpretable features aligned with Swiss-Prot[28] functional annotations. However, their approach fundamentally differs from ours by operating on per-residue token embeddings, enabling position-specific analysis within protein sequences. In contrast, molecular representations require handling entire chemical structures encoded as fixed-dimensional vectors, necessitating different validation strategies. Parsan *et al.*[29] extended SAE analysis to protein structure prediction through ESMFold[27], demonstrating steering capabilities for structural motifs. While their steering experiments parallel our molecular steering results, the token-level granularity again distinguishes their approach. The absence of an equivalent to Swiss-Prot annotations in chemistry – comprehensive, standardised functional labels with extensive literature evidence – required us to develop novel validation frameworks spanning multiple chemical abstraction levels.

S1.4 SMI-TED Architecture and Training

SMI-TED (SMILES-based Transformer Encoder-Decoder) is a 289M parameter transformer model that novelly combines molecular token encoding with SMILES reconstruction capabilities[7]. Unlike encoder-only models, SMI-TED employs a bidirectional transformer encoder (12 layers, 768 hidden dimensions, 12 attention heads) coupled with a decoder that reconstructs complete SMILES strings from learned representations.

The model processes SMILES through molecular tokenisation, decomposing chemical structures into substructure tokens from a vocabulary of 2,993 SMILES tokens. Each token embedding $\mathbf{x}_i \in \mathbb{R}^{768}$ passes through transformer layers that incorporate rotary position embeddings (RoFormer)[30], enabling better capture of molecular topology. Critically, SMI-TED introduces a novel submersion-immersion mechanism that maps token sequences to a unified molecular representation $\mathbf{z} \in \mathbb{R}^{768}$:

$$\mathbf{z} = \text{LayerNorm}\left(\text{GELU}(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\right)\mathbf{W}_2, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{L\times 768}$ represents the sequence of L token embeddings, $\mathbf{W}_1 \in \mathbb{R}^{L\times 768}$, $\mathbf{b}_1 \in \mathbb{R}^{768}$, and $\mathbf{W}_2 \in \mathbb{R}^{768\times 768}$. This latent representation enables both molecular property prediction and full SMILES reconstruction - a capability that enforces learning of complete chemical information.

S1.5 Sparse Autoencoders Formulation

An SAE can be generally defined by its encoder and decoder functions:

$$\left.\begin{aligned} \text{Encoder:} \quad & \mathbf{h} = f_{\text{enc}}(\mathbf{x}) \in \mathbb{R}^n \\ \text{Decoder:} \quad & \hat{\mathbf{x}} = f_{\text{dec}}(\mathbf{h}) \in \mathbb{R}^d \end{aligned}\right\} \text{SAE}(\mathbf{x}) = f_{\text{dec}}(f_{\text{enc}}(\mathbf{x})) = \hat{\mathbf{x}} \tag{2}$$

where $f_{\text{enc}} : \mathbb{R}^d \to \mathbb{R}^n$ maps the input $\mathbf{x}$ to a high-dimensional latent space ($n \gg d$) and $f_{\text{dec}} : \mathbb{R}^n \to \mathbb{R}^d$ reconstructs it. The model is trained by minimising a general loss function that balances reconstruction fidelity with a sparsity-inducing term[31]:

$$\mathcal{L}(\mathbf{x}) = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2}_{\text{Reconstruction}} + \underbrace{\lambda \mathcal{S}(\mathbf{h})}_{\text{Sparsity}} + \underbrace{\alpha \mathcal{L}_{\text{aux}}}_{\text{Auxiliary}} \tag{3}$$

This general formulation captures most SAE architectures through their specific definitions of the encoder f_{enc}, sparsity penalty $\mathcal{S}(\mathbf{h})$, and inclusion of an auxiliary loss.

In almost all cases, the decoder is a linear transformation $f_{\text{dec}}(\mathbf{h}) = \mathbf{W}_{\text{dec}}\mathbf{h} + \mathbf{b}_{\text{pre}}$, where $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{d \times n}$. The architectures differ primarily in their encoder and loss configuration.

Traditional L1 SAEs[10]: the encoder is $f_{\text{enc}}(\mathbf{x}) = \text{ReLU}(\mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{pre}}) + \mathbf{b}_{\text{enc}})$. Sparsity is encouraged via the L1-norm ($\mathcal{S}(\mathbf{h}) = \|\mathbf{h}\|_1$), and an auxiliary loss is generally not used ($\alpha = 0$).

TopK SAEs[13]: the encoder is $f_{\text{enc}}(\mathbf{x}) = \text{TopK}(\text{ReLU}(\mathbf{W}_{\text{enc}}(\mathbf{x} - \mathbf{b}_{\text{pre}}) + \mathbf{b}_{\text{enc}}), k)$, where TopK sets all but the k largest elements to zero. The explicit sparsity penalty is absent ($\lambda = 0$), and an auxiliary loss is often included ($\alpha > 0$) to encourage feature utilisation and prevent dead features.

Matryoshka SAEs[32]: the key innovation is in the reconstruction loss. Instead of a single term, the loss is a sum over multiple nested dictionaries of increasing size, $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$. For each size $m \in \mathcal{M}$, a partial reconstruction $\hat{\mathbf{x}}^{(m)}$ is computed using only the first m latents and the corresponding columns of the decoder matrix:

$$\hat{\mathbf{x}}^{(m)} = \mathbf{W}_{\text{dec},0:m}\mathbf{h}_{0:m} + \mathbf{b}_{\text{pre}} \tag{4}$$

The total reconstruction loss is the sum of the errors for each of these partial reconstructions:

$$\mathcal{L}_{\text{recon}} = \sum_{m \in \mathcal{M}} \|\mathbf{x} - \hat{\mathbf{x}}^{(m)}\|_2^2 \tag{5}$$

This objective forces earlier features to learn general concepts, while later features can specialise.

S2 Experimental Setup

S2.1 Data Curation

S2.1.1 PubChem Training Data

Following the exact curation procedure described by Soares *et al.* [14], we independently filtered over 122 million molecules from PubChem (July 2025) using their reported filtering process: molecular validity verification, canonicalisation, deduplication; additionally we performed desalting before deduplication. After filtering, the dataset contained approximately 91 million molecules, from which we uniformly sample 5 million molecules for SAE training. While we cannot guarantee identical overlap with SMI-TED's training data (as it was not publicly released), following the same curation procedure should yield a dataset with highly similar distributional properties.

S2.1.2 Evaluation Data

The MOSES, ChEMBL35 and MITOTOX datasets were prepared identically to the PubChem training data (see: Appendix S2.1.1). The final dataset sizes after preprocessing are provided in each section below.

MOSES MOSES[17] is a diverse representation of drug-like small molecule space that includes molecules optimised for drug development. We combine the MOSES test sets (N = 352,299) into a single evaluation dataset for assessing feature generalisation beyond the training distribution, providing a test of whether features learned on PubChem generalise to pharmaceutically relevant chemistry.

ChEMBL ChEMBL is a data source of literature validated functional and physicochemical annotations of molecules. In order to investigate functional relationships, we retrieved all small molecules ($<$ 500 Da) from ChEMBL35 [15]. Using the same preprocessing steps as the PubChem dataset, the resultant dataset contained 1,981,621 molecules. Calculated properties such as LogD, and functional measures such as binding affinity and targets were retrieved.

MITOTOX MITOTOX[16] is a dataset of small molecules with related mitochondrial toxicity annotations. A prepared subset was retrieved as per chemeleon-tox[33], and resulted in 3,742 molecules; 529 of which were labelled toxic (14.1%).

Physicochemical Descriptors and Substructures For all molecules Mordred 2D descriptors[34,35] (N = 1613) were calculated. For all molecules Atom Invariant Morgan fingerprints (radius = 2, use_chirality = True) were used to generate their substructure sets. For calculating the Tanimoto similarity distribution, a 4096-bit fingerprint size was used.

S2.2 SAE Training Details

We build atop the implementation of TopK SAEs by Samuel Marks and Mueller[36], originally developed for large language model interpretability. The official SMI-TED model is available on the Hugging Face Hub[37]. Each sweep configuration required approximately 3 GPU hours on an NVIDIA L4 GPU. During training, all molecular representations are normalised to have a unit mean squared norm as per [38].

Table S1: Hyperparameter configuration for TopK SAE Training. Values in parentheses represent the grid search range.

Hyperparameter	Value(s)
Dictionary Size Multiplier	(8, 16, 32)
Learning Rate (lr)	0.0001
Top-K (k)	(40, 80, 160)
AuxK Alpha (α)	0.03125
Training Epochs	80
Batch Size	256
Warmup Steps Fraction	0.05

The SAE training uses the Adam optimiser[39] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate schedule that includes linear warmup followed by linear decay beginning at 80% of total training steps.

The Warmup Steps Fraction parameter controls the proportion of total training steps during which the learning rate gradually increases from zero to its target value, implementing a learning rate warmup schedule that helps stabilise early SAE training and improve convergence.

The AuxK Alpha parameter controls the weighting coefficient for the auxiliary loss term in TopK SAE training, which encourages the model to use a broader set of features beyond just the top-k activations to improve feature diversity and reduce dead neurons.

S2.3 Evaluation Methodology

A core challenge in mechanistic interpretability is that standard SAE training metrics are only proxies for the true goal of discovering human-interpretable features, and developing robust interpretability metrics remains an open problem[24].

Our evaluation strategy is therefore twofold. First, we assess the SAE's reconstruction fidelity - its ability to preserve the essential chemical information from the original model. Second, we evaluate

the chemical meaningfulness of the learned features using a hierarchical framework designed to probe for specific, domain-relevant concepts.

S2.3.1 Reconstruction Fidelity Metrics

To measure how well the SAE's reconstructed vector, $\hat{x}$, preserves the information in the original vector, x, we use a combination of standard and domain-specific metrics.

Our primary measure is **functional fidelity**, defined as the success rate of decoding an SAE-reconstructed vector back into a chemically valid and equivalent SMILES string. This is a particularly stringent criterion, as minor errors can render a SMILES string invalid. We measure this at two levels: **strict accuracy** (exact canonical string matching) and **stereo accuracy** (chemical equivalence ignoring stereochemistry). High functional fidelity thus provides direct evidence that our SAE preserves the essential chemical information required by the foundation model. We supplement this with several standard metrics, which we define in Table S2.

Table S2: Standard metrics used to evaluate SAE reconstruction fidelity.

Metric & Description	Formula
L2 Reconstruction Loss The primary training objective.	$\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$
Fraction of Variance Explained Quantifies the variance of the original vector captured by the reconstruction.	$1 - \frac{\text{Var}(\mathbf{x}_i - \hat{\mathbf{x}}_i)}{\text{Var}(\mathbf{x}_i)}$
Fraction Alive The percentage of SAE features that activate on at least one molecule in the validation set.	-
Delta Loss Measures the preservation of the original model's loss landscape. $\mathcal{L}_{\text{SMI-TED}}$ is the original model's loss, composed of a token prediction cross-entropy term and an embedding MSE term[7]. A low $\Delta\mathcal{L}$ indicates high preservation.	$\Delta\mathcal{L} = \mathcal{L}_{\text{SMI-TED}}(\hat{\mathbf{x}}_i) - \mathcal{L}_{\text{SMI-TED}}(\mathbf{x}_i)$

S2.3.2 Framework for Evaluating Chemical Meaning

To systematically probe for chemical meaning, we validate features against hierarchical framework designed to capture the multi-scale nature of molecular properties. This framework consists of three categories: **substructural patterns**, which are local, discrete motifs such as functional groups and ring systems; (2) **physicochemical properties**, which are global, often continuous, properties emerging from the entire structure, like molecular weight or topological polar surface area, or systematic, high-dimensional features that encode topological, electronic, and geometrical information such as Mordred descriptors; and (3) **functional relationships**, which are abstract classifications, such as pharmacological drug class, that may not be apparent from simple structural similarity alone.

We use a logistic regression framework with a fixed random seed and class-balanced weights for two separate analyses. All models are trained and evaluated using a 5-fold cross-validation scheme.

Substructure Detection To evaluate how well individual features and neurons detect specific functional groups, we train a separate logistic regression model for each feature and each neuron. The model's task is to predict the presence or absence of a single functional group. We report the maximum F1 score achieved across the validation folds as the primary performance metric.

$$\left.\begin{aligned}\text{Precision} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \\ \text{Recall} &= \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}\end{aligned}\right\} \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Physicochemical We calculate the pairwise Spearman correlation coefficient between each feature- and neuron-descriptor pair, and apply identical analysis to PCA components and NMF factors (the latter requiring a shift transformation to ensure non-negativity) extracted using scikit-learn[40]. Where the set of descriptors is the Mordred 2D descriptors. Correlations with corresponding p-values < 0.05 were considered significant. The absolute value of the Spearman's ρ was used to rank the strength of the relationship.

Toxicity Prediction To assess the representational efficiency for a downstream task, we train two multiple logistic regression models to predict toxicity on the MITOTOX dataset. One model uses the complete set of SAE features as input, while the other uses all raw neuron activations. We identify inputs that are significantly predictive of toxicity by selecting model coefficients with a p-value < 0.05. The performance of each model is given by the area under the precision-recall curve (AUC_{pr}).

S3 Feature Characterisation

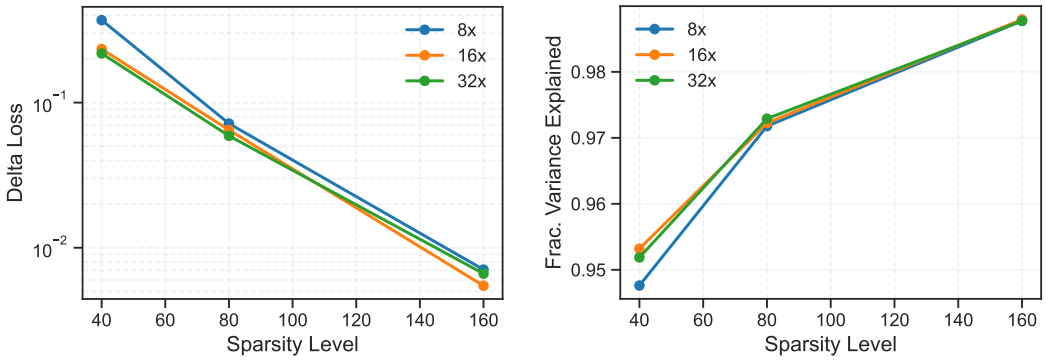
S3.1 Sparsity-Fidelity Trade-offs

To systematically evaluate SAE configurations, we analysed reconstruction quality and feature utilisation across a representative subset of 10,000 molecules uniformly sampled from the MOSES dataset (described in Appendix S2.1.2).

Figures S1–S3 characterise the trade-off space across expansion factors ($8\times$, $16\times$, $32\times$) and sparsity levels (40, 80, 160), revealing distinct operating regimes for each configuration.

The delta loss curves (Figure S1a) demonstrate an exponential relationship between sparsity and downstream task preservation. Interestingly, expansion factor shows minimal impact on delta loss at matched sparsity levels, suggesting that dictionary size primarily affects feature granularity rather than reconstruction quality.

The plot of fraction of variance explained (Figure S1b) shows that reconstruction fidelity increases with k, but with diminishing returns. This analysis informed our selection of $k = 80$ for the primary model, as it captures a high proportion of the original vector's variance (≈ 0.972) while a further doubling of k to 160 yields only a marginal improvement (to ≈ 0.987).



(a) Delta Loss vs. Sparsity Level. This plot shows the change in the foundation model's loss when using the SAE-reconstructed vector instead of the original. Lower values indicate better preservation of the original model's loss landscape. As expected, reconstruction fidelity improves (delta loss decreases) as the number of active features increases.

(b) Fraction of Variance Explained vs. Sparsity Level. This plot quantifies the proportion of the original activation vector's variance captured by the SAE's reconstruction at different sparsity levels. Higher values indicate a more faithful reconstruction of the original vector. As expected, reconstruction fidelity improves as the number of active features increases.

Figure S1: Sparsity-Fidelity Trade-off Across SAE Configurations.

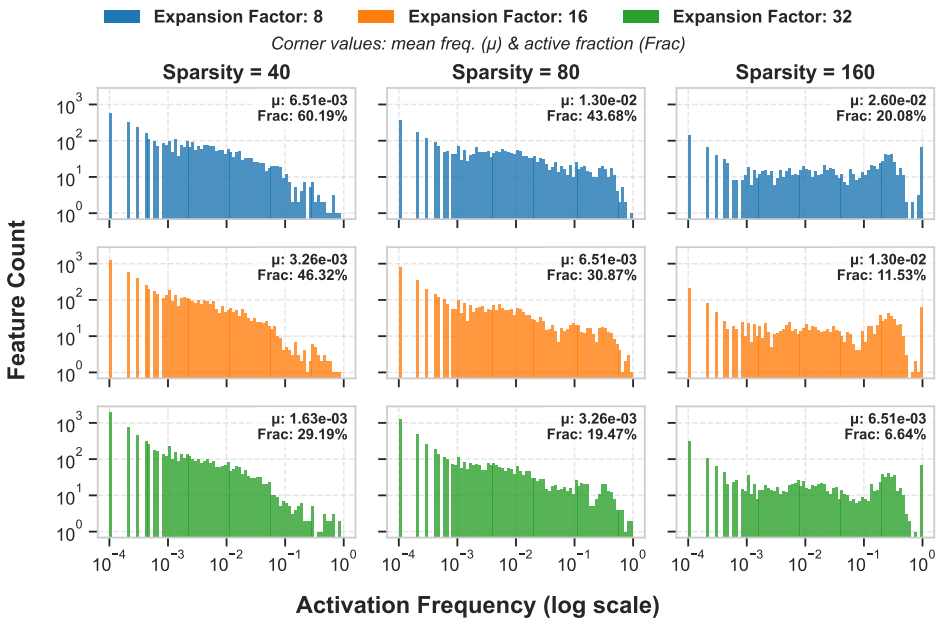
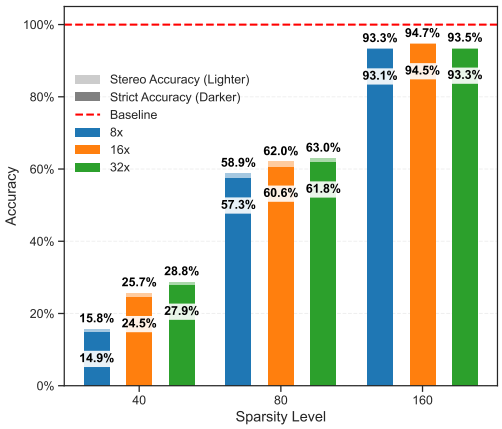
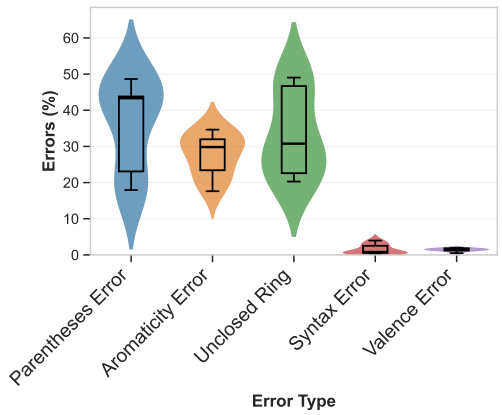


Figure S2: Feature activation frequency distributions across SAE hyperparameter configurations. Each subplot shows the histogram of activation frequencies for individual SAE features, organised by expansion factor (rows) and sparsity level (columns). Histograms use logarithmic binning and scaling to visualise the characteristic heavy-tailed distribution of feature activations.

S3.2 SMILES Reconstruction Analysis



(a) Comparison of SMILES reconstruction accuracy across different SAE hyperparameter configurations. Dark bars show strict accuracy (exact string matching), while light bars show stereo accuracy (chemical equivalence ignoring stereochemistry). The red dashed line indicates baseline model performance without SAE reconstruction. Results demonstrate how reconstruction accuracy varies with SAE architecture parameters, with higher expansion factors generally maintaining better accuracy recovery across sparsity levels.



(b) Distribution of SMILES reconstruction error types across SAE hyperparameter configurations. Box plots show the percentage distribution of different error categories (valence errors, aromaticity errors, bond duplication, unclosed rings, parentheses errors, and syntax errors) that occur when SAE-reconstructed embeddings are decoded back to SMILES strings. Error categories are classified using standardised RDKit parsing error analysis to understand how SAE reconstruction affects different aspects of molecular structure representation.

Figure S3: Impact of SAE Hyperparameters on SMILES Reconstruction Fidelity.

S4 Supplementary Results

S4.1 Substructures

A total of 14 common functional groups[41] were retrieved in SMARTS format to provide a range of substructures present in the dataset at varying prevalence. The SMARTS strings were used to identify the presence/absence of the functional group. Some molecules contain the same functional group multiple times, or multiple functional groups.

Steering was performed by intervening on the specified feature values, and setting the feature activation to 0.

Table S3: Maximum F1 scores and prevalence in 5M PubChem for various functional groups

Functional Group	**Maximum F1**		**Prevalence (%)**
	Features	**Neurons**	
Alkyl Carbon	**0.945**	0.938	88.286
Acetylenic Carbon	**0.933**	0.079	0.758
Carbonyl group, High specificity	**0.745**	0.735	51.197
Cyanamide	**0.667**	0.030	0.086
Ether	**0.792**	0.655	36.245
Primary amine, not amide	**0.697**	0.431	8.613
Azo nitrogen	**0.706**	0.071	0.637
Nitrate	**1.000**	0.056	0.103
Hydroxyl	**0.838**	0.654	39.569
Peroxide groups	**0.667**	0.043	0.224
Phosphoric acid group 1	**0.571**	0.075	0.448
Thiol	**0.900**	0.135	0.982
Sulfide	**0.700**	0.344	7.907
Chloride (Carbon-attached)	**0.802**	0.533	21.258

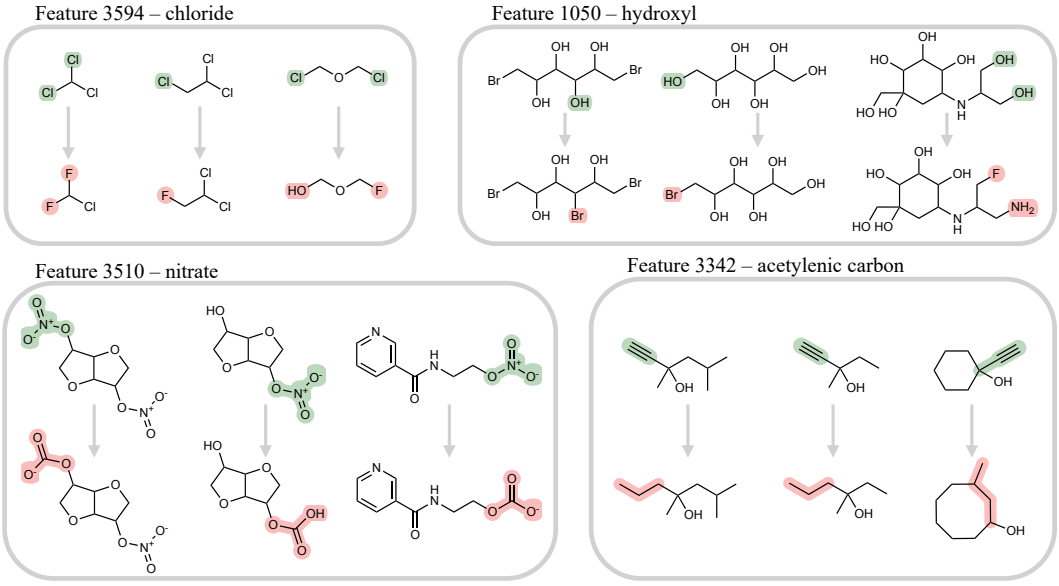


Figure S4: Top three activated molecules for the top correlated feature for functional groups chloride, hydroxyl, nitrate, and acetylenic carbon. Molecules are then steered by setting the specified feature activation to 0, and the altered substructure is highlighted before (green) and after (red) steering.

S4.2 Physicochemical Properties

The top 3 feature-descriptor relationships, ranked by Spearman correlation, were selected. The top 3 molecules which had the highest activations for each corresponding feature were retrieved. The feature activation was set to zero, and the corresponding descriptor was recalculated. These were (with Spearman's ρ): StsC; sum of tsC ($\rho = 0.89$), SMR_VSA7; MOE MR VSA Descriptor 7 ($3.05 \leq x < 3.63$) ($\rho = 0.85$)and Xch-3d; 3-ordered Chi chain weighted by sigma electrons ($\rho = 0.75$).

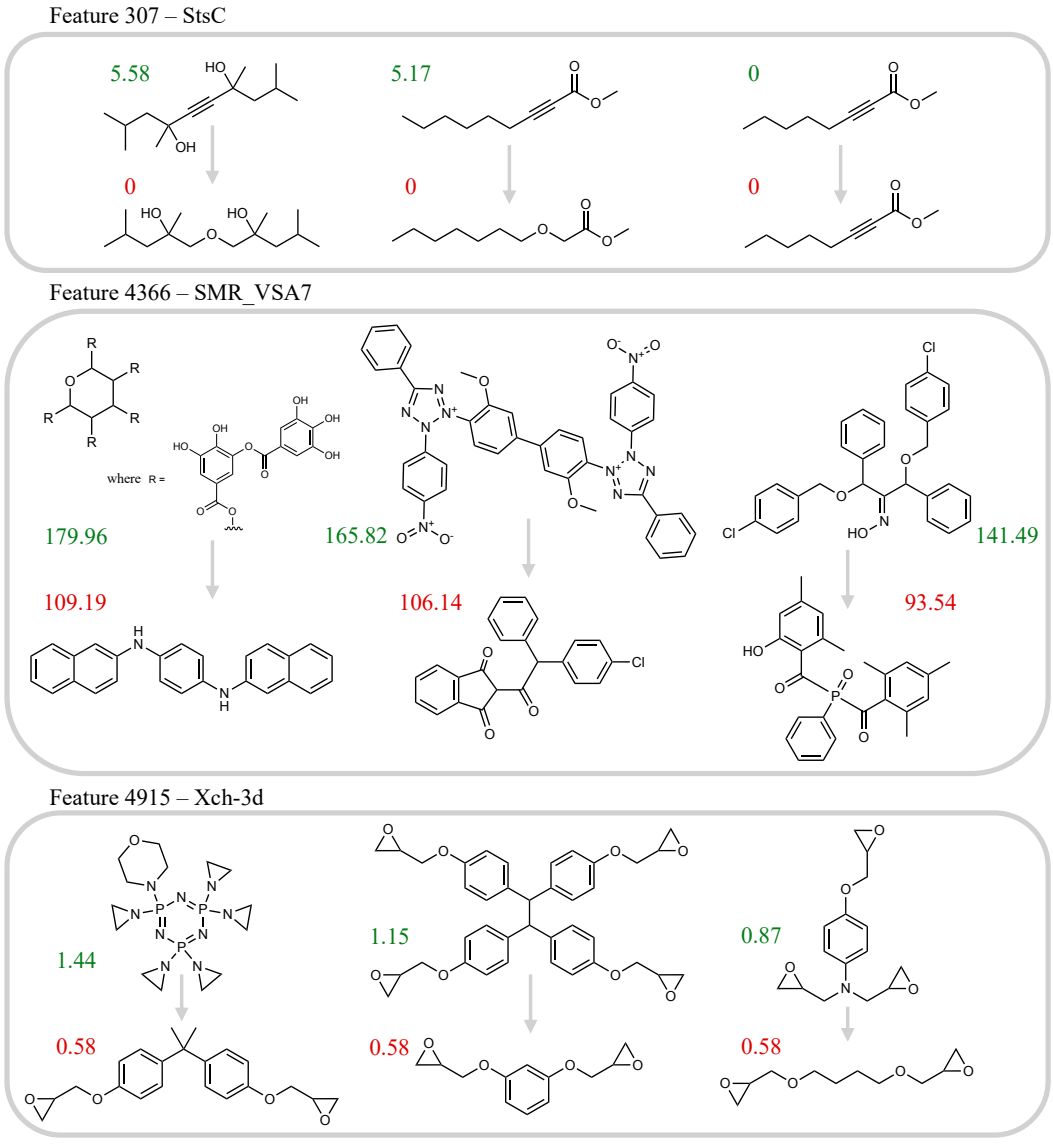


Figure S5: Top three activated molecules for the top three feature-descriptor relationships: StsC, SMR_VSA7, Xch-3d. Molecules are then steered by setting the specified feature activation to 0, and the descriptor value is provided before (green) and after (red).

S4.3 Functional Behaviour

ChEMBL35 was subset to molecules with at least one Ki - ChEMBL target pair. For each feature, the molecules for which the feature had a normalised activation > 0.5 were selected. The set intersection of ChEMBL targets for this selection was calculated. Then the largest set was retrieved.

To establish a null distribution for assessing Tanimoto similarity (using radius-2, 4096-bit Feature-AtomInv Morgan fingerprints, with chirality), we first estimate the required sample size of random molecular pairs.

Given the equation for estimating a proportion (or similarity) with a given margin of error ϵ at a confidence level of z:

$$m = \frac{z^2 \sigma^2}{\epsilon^2} \tag{7}$$

Using a pilot study variance estimate of $\hat{\sigma} = 0.062$, from 100,000 pairs a desired margin of error of $\epsilon = 0.0001$, and a 99% confidence level, we calculate a required sample size of approximately 2.6 million pairs. We subsequently sample 10 million random molecular pairs from the ChEMBL dataset to construct a high-resolution empirical null distribution of Tanimoto similarity scores. This distribution serves as the baseline for computing the statistical significance of observed similarity values (see Figure S6).

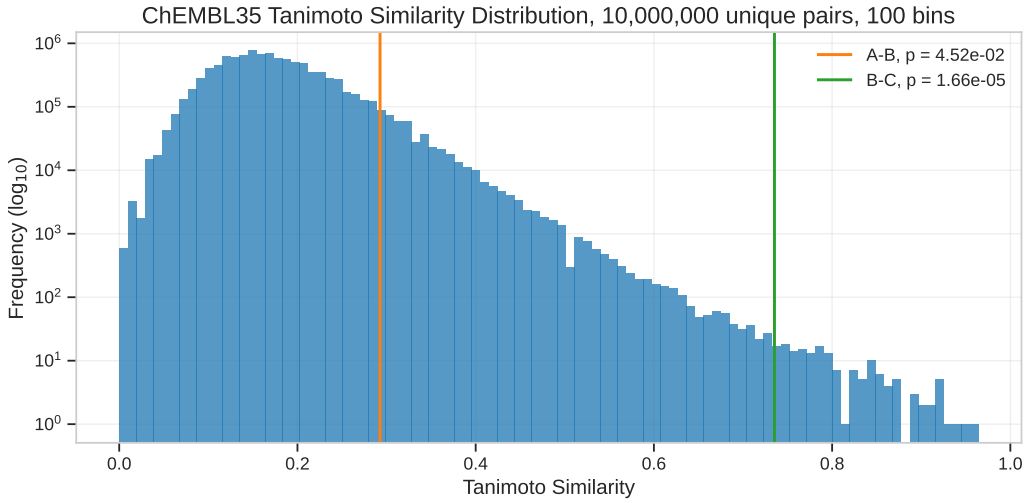


Figure S6: Distribution of Tanimoto Similarity (TS) across 10,000,000 unique pairs, sampled from ChEMBL35. The TS of CHEMBL1672485 (A) and CHEMBL454618 (B) (A-B: 0.293) is shown in orange, the TS of B and CHEMBL368522 (C) (B-C: 0.735) is shown in green. The corresponding p-values drawn from a right-tailed empirical distribution.

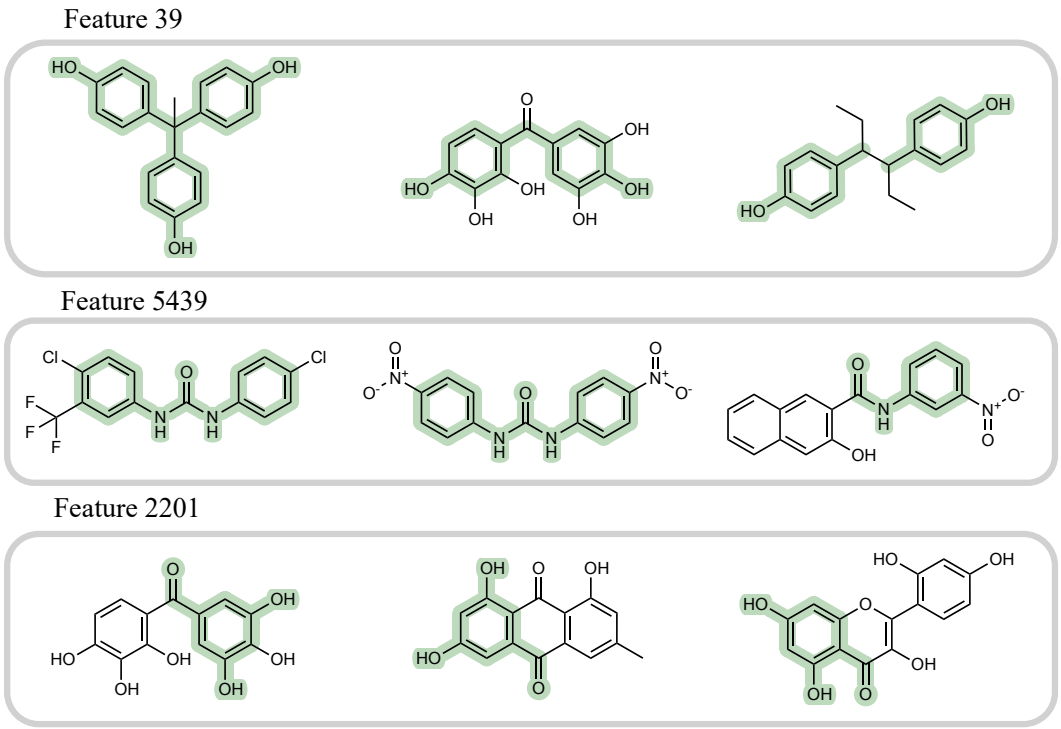


Figure S7: Top three activated molecules for the top three features used for logisitic regression of toxicity (see Section 3.3). The major common substructure for each feature is highlighted in green.

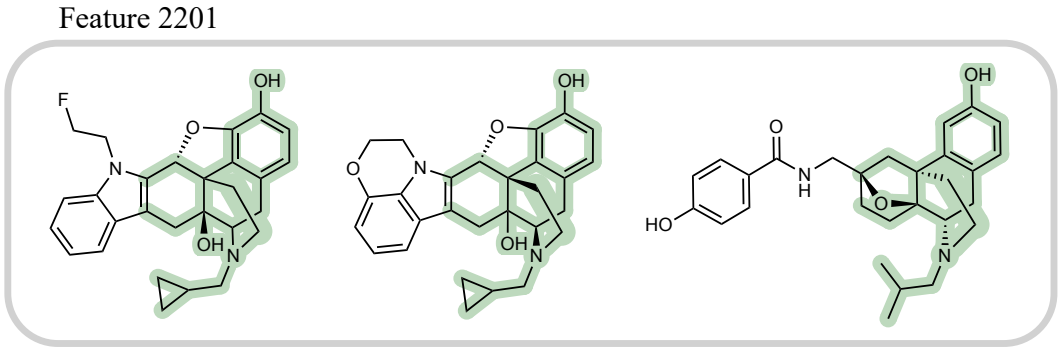


Figure S8: Top three activated molecules for the feature found to be related to pharmacological function to opioid receptors (see Section 3.3). The major common substructure is highlighted in green.

S5 Steering Stability

To evaluate the stability and causal influence of the learned representations, we conduct a series of ablation experiments comparing the original SMI-TED neuron activations (dense representations) with our SAE-derived features (sparse representations). The experiments are performed on a 10,000-molecule subset of the MOSES dataset.

For the dense neuron representations, we intervene on each of the 768 neurons individually. For each neuron, we identify the 100 molecules that elicit its highest and lowest activations. Assuming a normal distribution of activations for a given neuron, we ablate its value to the distribution's mean for these selected molecules before decoding them back to SMILES strings. This intervention results in minimal change: only 14 of the 768 neurons produce any invalid SMILES upon ablation, and no

interventions result in a valid but different molecule. This suggests the dense representations are highly robust, with chemical information distributed across many neurons, making individual neurons non-critical for reconstruction.

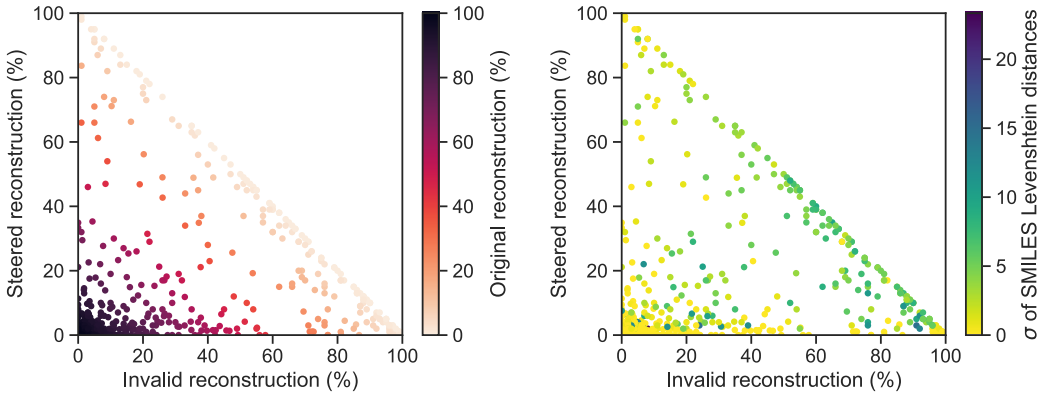


Figure S9: Reconstruction rates of 100 molecules after ablation of 2501 active features on a 10k subset of MOSES. After ablation of a feature, embeddings are decoded into either the original SMILES ("Original", hue (left)), a different SMILES ("Steered", y-axis), or an invalid SMILES ("Invalid", x-axis). The variability in steered transformations are shown by the standard deviations of Levenshtein distances between the original and steered SMILES strings (hue, right).

For the sparse SAE features, we perform a different intervention tailored to their sparse nature. For each of the 2,501 active features, we select up to 100 molecules where that feature has a non-zero activation. We then ablate the feature by setting its activation to zero, effectively "turning it off," before decoding. The outcomes are then categorised as: 1) *Original*: the decoded SMILES matches the original, 2) *Invalid*: the decoded SMILES is chemically invalid, or 3) *Steered*: the decoded SMILES is valid but different from the original.

In contrast to the dense neurons, the sparse features demonstrate significant steerability. We find that interventions on 749 of the 2,501 active features successfully steer molecules to new, valid chemical structures. As shown in Figure S9, this approach reveals a clear trade-off between feature stability (valid reconstruction) and steerability, confirming that individual SAE features often represent specific, manipulable chemical concepts. Levenshtein distances between the original and steered SMILES strings are also used as an approximation for measuring the consistency of steering transformations for a given feature[42]. The Levenshtein distance is a metric for measuring the difference between two string sequences. As expected, the most stable and steerable features have a low standard deviation of Levenshtein distances, indicating that steering most likely changes most molecules in the same way. Features with a higher invalid reconstruction rate also have a higher standard deviation, indicating that their steered changes are less consistent.