Interpreting GFlowNets for Drug Discovery: Extracting Actionable Insights for Medicinal Chemistry

Amirtha Varshini A S
Montai Therapeutics
asindhanai@montai.com

Duminda S. Ranasinghe
Montai Therapeutics
dranasinghe@montai.com

Hok Hei Tam
Montai Therapeutics
htam@montai.com

Abstract

Generative Flow Networks (GFlowNets) provide a powerful framework for molecular design, yet their internal decision policies remain opaque. This hinders adoption in drug discovery, where chemists require interpretable rationales for generated molecules. We introduce a unified interpretability framework for hierarchical GFlowNets applied to reaction graphs. Our approach integrates: (i) gradient-based saliency with counterfactual perturbations, (ii) concept attribution via sparse autoencoders (SAEs), and (iii) motif probes for structural motif recovery. Applied to SynFlowNet [1], our results suggest that drug-likeness (QED) can be decomposed into interpretable axes such as polarity and lipophilicity, while probes recover motifs including halogens and ring systems. Counterfactual saliency (cFS) further identifies substructures whose modification directly alters predicted rewards, offering actionable, causally grounded interpretability. Together, these results extend interpretability to structured generative models and offer insights for medicinal chemistry.

1 Introduction

Drug discovery requires navigating vast chemical spaces under strict synthesizability and drug-likeness constraints. Deep generative models such as VAEs, GANs, and diffusion models have shown promise in molecule generation, but their black-box nature limits adoption in practice.

GFlowNets [3, 4] provide an alternative: they model stochastic policies that construct molecules sequentially. SynFlowNet [1], a hierarchical GFlowNet, assembles molecules via reaction templates and building blocks. Despite strong generative performance, interpretability of GFlowNets remains unexplored.

We contribute a toolkit that adapts interpretability methods from supervised learning [5–7] to structured generative settings. Our framework bridges machine learning explanations with medicinal chemistry reasoning, aligning with recent calls for explainability in drug discovery [8, 9].

2 Methods

In our experiments, we analyze SynFlowNet trained with QED (Quantitative Estimate of Drug-likeness) as the reward function [1]. Our interpretability framework combines three complementary approaches, each targeting a different level of explanation. Gradient-based saliency highlights which atoms and bonds drive specific generative actions, providing fine-grained local attribution. Counterfactual perturbations extend this by testing causal influence of structural edits on action probabilities and reward outcomes [9, 10]. Sparse autoencoders (SAEs) uncover disentangled, axis-aligned latent factors, enabling analysis of how abstract representations align with physicochemical properties [6, 11]. Finally, motif probes test whether discrete chemical motifs are encoded in embeddings, linking model internals to recognizable medicinal chemistry concepts. Together, these methods provide multi-scale interpretability, spanning from atom-level rationales to latent-space structure and motif-level detectors.

Massachusetts Institute of Technology - Molecular Machine Learning Conference 2025.

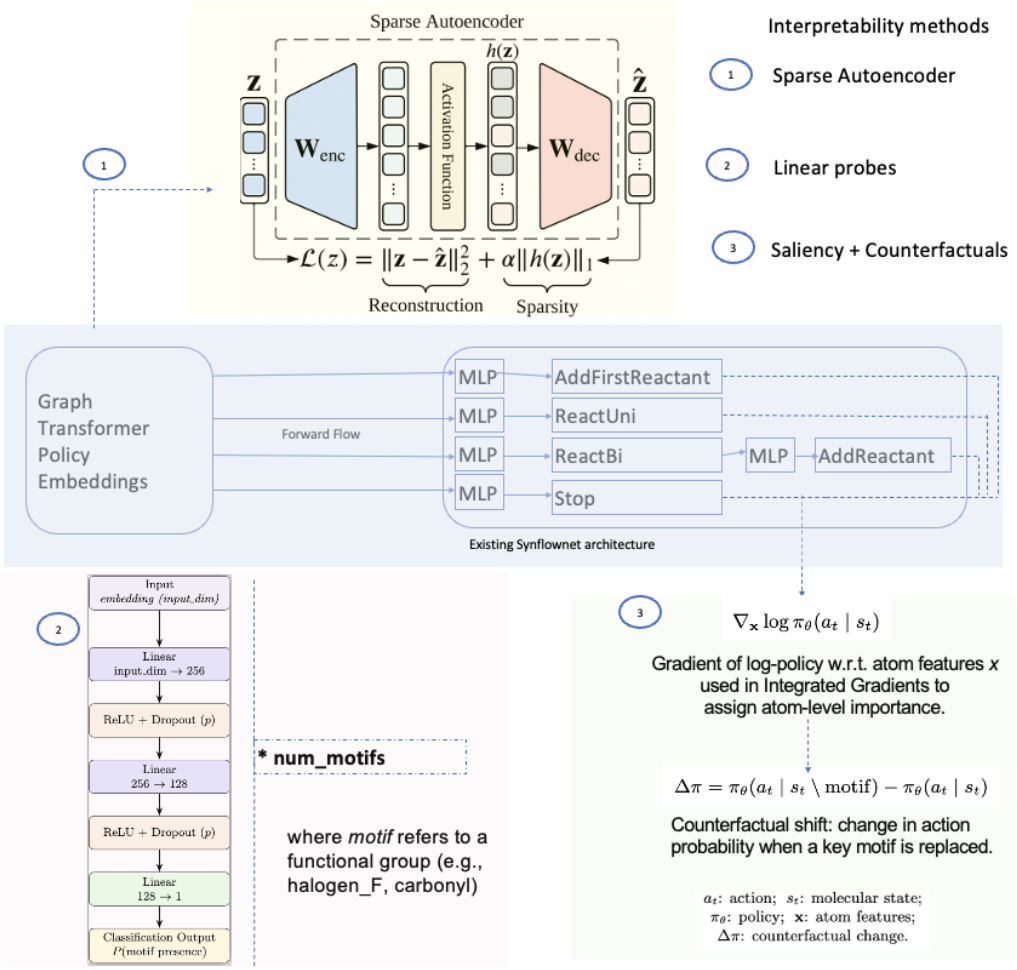

Figure 1: Overview of the proposed interpretability framework for GFlowNets. Our pipeline integrates (1) sparse autoencoders [2](SAEs) for discovering disentangled chemical factors such as polarity and lipophilicity, (2) motif probes to test whether embeddings encode functional groups, and (3) gradient-based saliency and counterfactual perturbations for atom- and motif-level attribution. Together, these approaches span fine-grained atomic rationales to high-level medicinal chemistry concepts.

2.1 Gradient-Based Saliency with Counterfactuals

We compute per-atom saliency by backpropagating through action log-probabilities, following attribution methods such as integrated gradients [5]. Specifically, for trajectory action a_t with probability $\pi_\theta(a_t|s_t)$, we compute $\nabla_x \log \pi_\theta(a_t|s_t)$, where x indexes atom features. This highlights atoms most responsible for the chosen action.

To move beyond correlation, we explore counterfactual perturbations. Using SMARTS-based substructure masking, we replace or delete motifs and recompute action probabilities. The relative shift $\Delta\pi$ provides exploratory evidence of causal influence:

$$\Delta\pi = \pi_\theta(a_t|s_t \setminus \text{motif}) - \pi_\theta(a_t|s_t).$$

To summarize the steps used in the current counterfactual analysis,

1. Infer the `Stop` action index robustly (using context if available, or the final logit otherwise);
2. Run Integrated Gradients (IG) on $\log p_\theta(\text{Stop})$ to obtain atom-level saliency scores;

3. Extract the top connected "hot" motifs corresponding to the most salient atoms; and
4. Apply a small set of safe `RDKit` edits (substitutions, deletions, or neutral replacements) within those motifs and report the change in reward ΔQED.

2.2 Sparse Autoencoders (SAEs)

To analyze whether SynFlowNet embeddings capture chemically meaningful structure, we apply sparse autoencoders (SAEs). Given a hidden embedding $h \in \mathbb{R}^d$ from the policy network, the SAE encodes a nonnegative, sparse code $z = \mathrm{ReLU}(Wh + b)$ and reconstructs $\hat{h} = W'z + b'$. Training minimizes

$$\mathcal{L} = \|h - \hat{h}\|_2^2 + \lambda \|z\|_1,$$

where the L_1 term encourages sparse, axis-aligned factors. For comparison, we also support a KL sparsity penalty on pre-activation logits to target mean activity ρ. In practice, we correlate each factor with molecular descriptors (e.g., TPSA for polarity, Crippen logP for lipophilicity) to test whether these abstract dimensions map to interpretable physicochemical axes.

2.3 Motif Probes

Whereas SAEs identify continuous latent factors, motif probes test whether discrete chemical motifs are encoded in SynFlowNet embeddings. We freeze the pretrained GFlowNet and train shallow feed-forward classifiers (two-layer MLPs) on the embeddings h to predict motif presence. Labels are obtained automatically using RDKit SMARTS pattern matching for functional groups, aromatic rings, and halogens. High probe performance indicates that motif information is readily accessible in the embeddings, linking abstract representations back to recognizable medicinal chemistry concepts.

3 Results

QED disentanglement. SAEs trained on SynFlowNet embeddings uncover 128 latent factors with a mean sparsity of 0.105 (fraction of molecules activating each factor). Several factors show strong correlations with physicochemical properties: Factor_11 with size ($r = 0.76$), Factor_86 with polarity ($r = -0.57$), and Factor_118 with polarity ($r = 0.54$). Predictors trained on these factors achieve high R^2 on polarity (0.92) and size (0.71), outperforming direct prediction of composite QED (0.25). These results indicate that QED decomposes into more linearly predictable components such as polarity, lipophilicity, and size.

Motif detection. Motif probes achieve strong classification performance across functional groups. In particular, halogen substitutions, aromatic rings, and carbonyl groups are predicted with AUROC scores exceeding 0.9 (see Supplementary Table 3). This suggests that SynFlowNet embeddings readily encode chemically meaningful motifs, consistent with medicinal chemistry heuristics, even when probes are limited to shallow classifiers.

Counterfactual analysis. The counterfactual-saliency module highlights reward-sensitive regions by integrating atom-level gradients and localized structure edits. For representative trajectories (Fig. 2B), the most salient motifs correspond to polar substituents and aromatic fragments; safe edits within these motifs produce consistent ΔQED shifts aligned with chemical expectations. These results confirm that counterfactual perturbations provide stable, interpretable signals of causal influence on reward outcomes.

4 Discussion

Our results demonstrate that interpretability methods such as gradients, saliency and counterfactuals, and concept attribution, can be adapted to structured generative models. The combined saliency and counterfactual framework (cFS) provides mechanistic insight into how specific atomic environments modulate action probabilities and downstream rewards. These explanations align naturally with medicinal chemistry reasoning—capturing polarity, lipophilicity, aromaticity, and halogenation—and support transparent, design-relevant interpretation of GFlowNet policies.

At the same time, several limitations remain. Our study focuses largely on QED; extending to multi-objective settings (e.g., synthetic accessibility, binding affinity) is necessary to assess generalization.

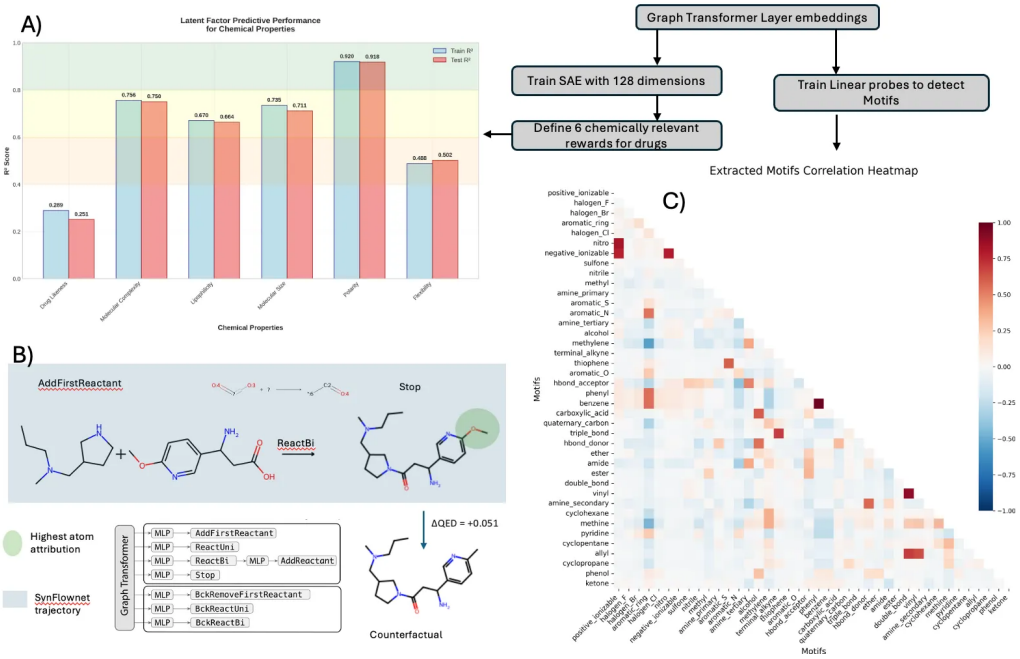

Figure 2: Interpretability results on SynFlowNet embeddings. (A) Predictive performance of sparse autoencoder (SAE) factors across six chemical properties, showing that factors disentangle polarity, size, and lipophilicity more effectively than composite QED. (B) Example SynFlowNet trajectory with atom-level saliency (highlighted atoms) and a counterfactual edit that alters predicted QED, illustrating causally grounded attribution. (C) Motif–factor correlation heatmap from motif probes, revealing that embeddings encode functional groups such as halogens, aromatic rings, and carbonyl groups with high fidelity.

Second, our disentanglement analysis relies on sparse autoencoders and motif probes, which assume relatively simple structure in latent spaces. Although this provides an accessible first step, it remains unclear whether more complex, non-linear objectives would admit similar decomposition.

In line with recent discussions [1], one intriguing direction is to investigate whether conditioning GFlowNets directly on physicochemical properties, rather than post hoc disentanglement, could yield latent representations more naturally aligned with domain-relevant axes.

5 Conclusion

We introduced an interpretability framework for GFlowNets in molecular design, integrating saliency, counterfactuals, and latent factor attribution. This framework reveals chemically meaningful motifs and properties within SynFlowNet's embeddings, bridging machine learning interpretability and chemical reasoning.

Our results highlight both the promise and the limitations of current approaches. By showing that GFlowNets encode interpretable factors such as polarity and lipophilicity, we take a step toward actionable insights for drug discovery. However, broader evaluation—including quantitative impact on molecule design, comparisons with alternative explainers, and application to multi-objective scenarios—remains essential. Future work should explore conditioning GFlowNets directly on physicochemical properties to induce representations more closely aligned with medicinal chemistry.

Overall, this study provides an initial foundation for interpretable generative modeling in drug discovery, offering chemists a clearer window into how structured generative policies operate and how they may be trusted in practice.

Acknowledgements

We thank colleagues at Montai for feedback. We also acknowledge the use of OpenAI's ChatGPT in helping edit and refine the manuscript text.

References

[1] Miruna Cretu, Charles Harris, Ilia Igashov, Arne Schneuing, Marwin Segler, Bruno Correia, Julien Roy, Emmanuel Bengio, and Pietro Liò. Synflownet: Design of diverse and novel molecules with synthesis constraints. *arXiv preprint arXiv:2405.12345*, 2024.

[2] Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *EMNLP 2025 Findings*, 2025.

[3] Yoshua Bengio, Tristan Deleu, Edward Hu, Salem Lahlou, Minka Li, and Emmanuel Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Kanika Madan, Moksh Jain, and Yoshua Bengio. Learning gflownets from partial episodes for improved exploration in compositional spaces. In *International Conference on Machine Learning (ICML)*, 2022.

[5] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[6] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

[9] Geemi P Wellawatte, Arvind Seshadri, and Andrew D White. Counterfactual explanations for molecules. *Machine Learning: Science and Technology*, 3(4):045009, 2022.

[10] Ana Lucic, Hinda Haned, Jefrey Lijffijt, and Tijl De Bie. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[11] Róbert Csordás, Alex Wang, and Jürgen Schmidhuber. Sparse autoencoders for interpretability: disentangling latent space in transformers. *arXiv preprint arXiv:2309.XXXX*, 2023.

A Code Availability

The full implementation of our interpretability framework, including sparse autoencoders, motif probes, and counterfactual analysis, is available at:
https://github.com/amirtha-montai/synflownet_public/tree/main/src/interpretability

B Sparse Autoencoder Chemical Factor Analysis

To better understand how QED (drug-likeness) is represented internally, we trained a sparse autoencoder (SAE) on SynFlowNet embeddings. The SAE disentangled QED into interpretable chemical axes such as *size*, *polarity*, and *lipophilicity*, showing that drug-likeness emerges from simpler physicochemical components rather than a single latent dimension.

B.1 Dataset Summary

- Total molecules analyzed: 32,054

- Embedding dimension: 256
- Number of latent factors discovered: 128
- Number of reward signals: 6
- Train/test split: 28,848 / 3,206

B.2 Reward Prediction Performance

Reward Signal	Train R^2	Test R^2
Drug-likeness	0.289	0.251
Complexity	0.756	0.750
Lipophilicity	0.670	0.664
Size	0.735	0.711
Polarity	0.920	0.918
Flexibility	0.488	0.502

Table 1: Predictive R^2 scores for six chemical reward signals using sparse autoencoder latent factors. Polarity and size are well captured ($R^2 > 0.7$), whereas composite drug-likeness (QED) is harder to predict directly, suggesting that interpretable components underpin QED.

B.3 Latent Factor Sparsity

Sparsity statistics (fraction of molecules with factor activation > 0.1):

- Mean: 0.105
- Std: 0.171
- Min: 0.000
- Max: 0.666

B.4 Top Factor–Reward Associations

Factor	Reward Signal	Correlation (r)
Factor_11	Size	0.757
Factor_75	Size	-0.574
Factor_86	Polarity	-0.570
Factor_118	Polarity	0.540
Factor_28	Size	0.525
Factor_12	Size	-0.507
Factor_52	Polarity	0.446
Factor_49	Size	-0.422
Factor_34	Polarity	-0.415
Factor_96	Complexity	-0.412

Table 2: Latent factors extracted by sparse autoencoders and their strongest correlations with chemical reward signals. Several factors align with interpretable physicochemical properties such as size (Factor_11) and polarity (Factors_86, Factor_118).

B.5 Reward-Specific Factor Summary

- Drug-likeness: Factor_28 (0.379), Factor_62 (0.325), Factor_11 (0.310)
- Complexity: Factor_96 (0.412), Factor_29 (0.380), Factor_86 (0.365)
- Lipophilicity: Factor_11 (0.412), Factor_86 (0.387), Factor_118 (0.355)
- Size: Factor_11 (0.757), Factor_75 (0.574), Factor_28 (0.525)
- Polarity: Factor_86 (0.570), Factor_118 (0.540), Factor_52 (0.446)
- Flexibility: Factor_55 (0.375), Factor_11 (0.334), Factor_36 (0.300)

C Additional Motif Probe Results

Table 3 summarizes AUROC and average precision (AP) scores for motif classification using motif probes across all tested motifs. These results confirm that functional group information is readily accessible in the embeddings.

Table 3: Motif probe classification results across diverse functional groups. High *AUROC* (> 0.9) for halogens, aromatic rings, and ionizable groups demonstrates that SynFlowNet embeddings encode chemically meaningful motifs accessible to shallow classifiers.

Motif	Prevalence	AUROC	AP
positive_ionizable	0.0226	1.0000	1.0000
halogen_F	0.3222	1.0000	1.0000
halogen_Br	0.1229	1.0000	1.0000
aromatic_ring	0.9061	1.0000	1.0000
halogen_Cl	0.1467	1.0000	1.0000
nitro	0.0208	0.99998	0.99927
negative_ionizable	0.0221	0.99998	0.99924
sulfone	0.0484	0.99941	0.98622
nitrile	0.2253	0.99660	0.98899
methyl	0.7334	0.99588	0.99833
amine_primary	0.1375	0.99146	0.95530
aromatic_S	0.0658	0.99069	0.87042
aromatic_N	0.6129	0.98763	0.99259
amine_tertiary	0.4640	0.98726	0.98557
alcohol	0.4387	0.98442	0.98150
methylene	0.9379	0.97621	0.99842
terminal_alkyne	0.0258	0.97523	0.70965
thiophene	0.0296	0.97522	0.48240
aromatic_O	0.0878	0.97386	0.83362
hbond_acceptor	0.9326	0.97217	0.99793
phenyl	0.5976	0.97210	0.97911
benzene	0.5976	0.97199	0.97888
carboxylic_acid	0.2296	0.97099	0.90775
quaternary_carbon	0.1549	0.95900	0.83732
triple_bond	0.0422	0.95524	0.65181
hbond_donor	0.7903	0.94051	0.98364
ether	0.4434	0.93137	0.91106
amide	0.5278	0.92347	0.92691
ester	0.1167	0.92204	0.60984
double_bond	0.0610	0.92153	0.56933
vinyl	0.0559	0.92028	0.53539
amine_secondary	0.4357	0.91477	0.89198
cyclohexane	0.1252	0.90263	0.57151
methine	0.7588	0.89854	0.96505
pyridine	0.2656	0.88299	0.73224
cyclopentane	0.0547	0.88028	0.36046
allyl	0.0384	0.86442	0.33325
cyclopropane	0.1104	0.85982	0.42632
phenol	0.0499	0.85573	0.25594
ketone	0.0604	0.82974	0.36870