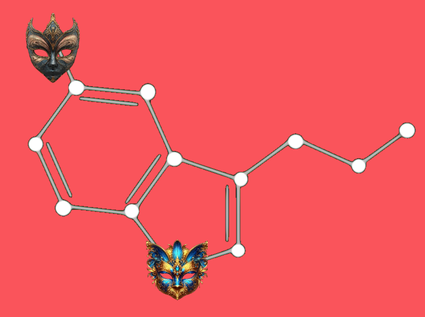
ENHANCING CHEMICAL EXPLAINABILITY THROUGH COUNTERFACTUAL MASKING

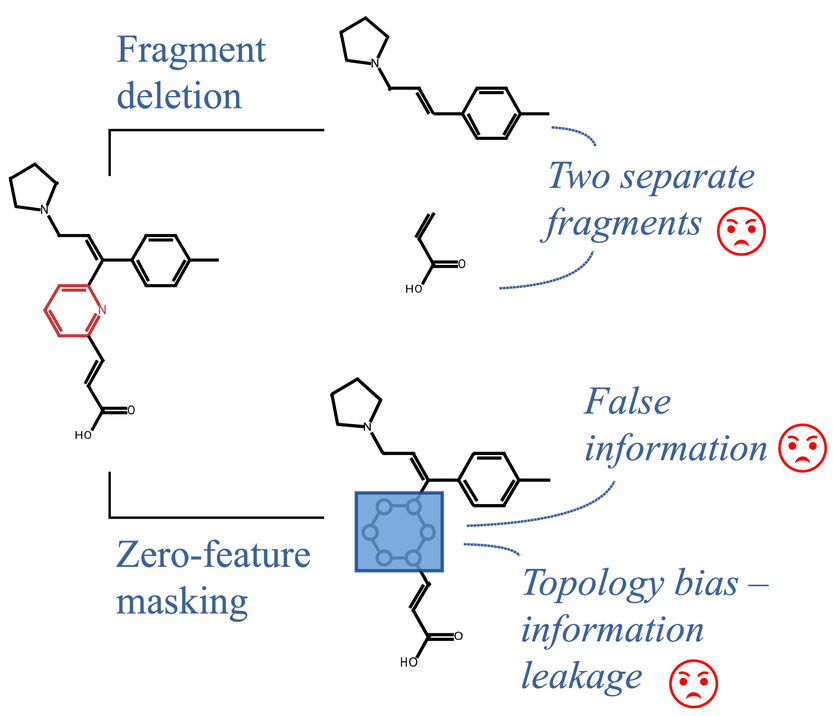
AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20–27, 2026 | SINGAPORE



Łukasz Janisiów Marek Kochańczyk Bartosz Zieliński Tomasz Danel
Jagiellonian University, Poland

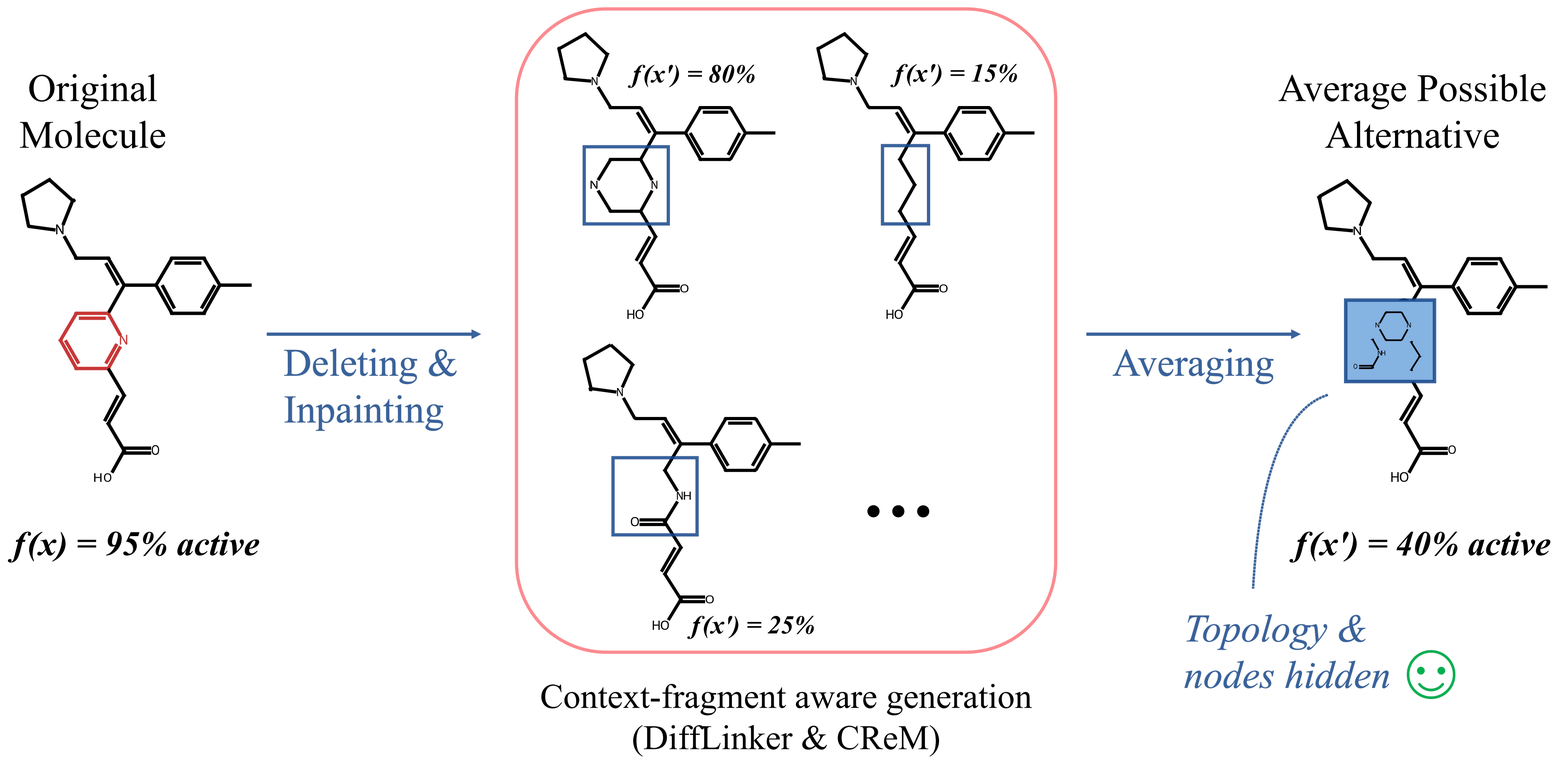
MOTIVATION: Molecular Masking Is Broken

When atoms or bonds are naively masked (by zeroing features or deleting parts of molecule), the resulting molecules often become **chemically implausible or physically impossible**, producing examples that fall outside the training distribution. This **out-of-distribution problem** undermines the reliability of both the explanations and their evaluation metrics.



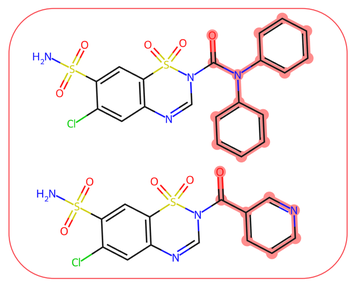
SOLUTION: Counterfactual Masking (CM)

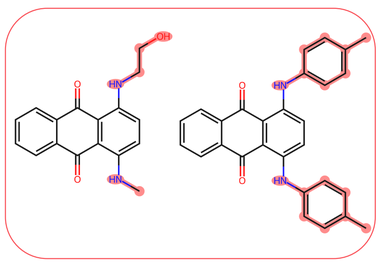
Counterfactual Masking masks by replacing the important molecular fragments with chemically plausible alternatives. This enables a direct comparison between the model's prediction for the original molecule and the average prediction across valid counterfactuals (i.e., all possible alternatives).

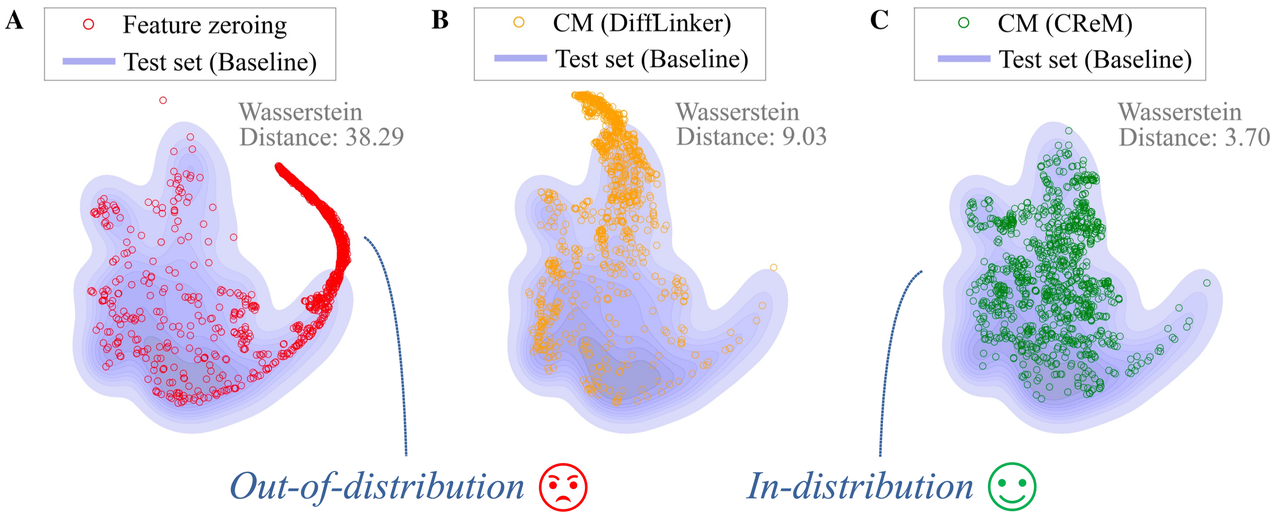


Evaluation of Fragment Masking

To evaluate different masking techniques, we used the common substructure pair dataset, which contains pairs of molecules with the same core structure. In each pair, we masked the fragments that were not shared and measured the difference in the model's predictions. An ideal masking technique should completely obscure the masked fragment, resulting in identical predictions for both molecules in the pair.



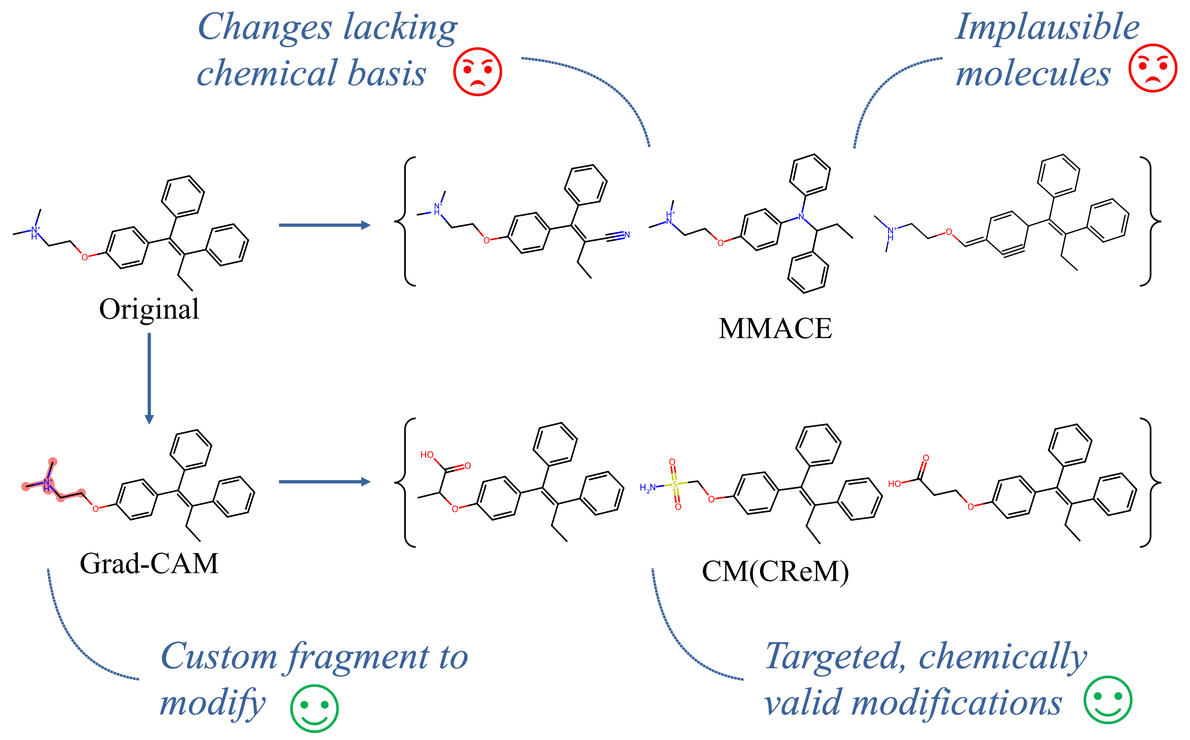




Masking	Single Anchor $\lvert\Delta\hat{y}\rvert\downarrow$	Single Anchor Validity ↑	Single Anchor Size ↑	Multiple Anchors $\lvert\Delta\hat{y}\rvert\downarrow$	Multiple Anchors Validity ↑	Multiple Anchors Size ↑	Both $\lvert\Delta\hat{y}\rvert\downarrow$	Both Validity ↑	Both Size ↑
No masking	2.81	100%	517	2.52	100%	638	2.91	100%	783
Feature zeroing	2.08	100%	517	1.60	100%	638	1.80	100%	783
CM (CReM)	1.68	88%	454	1.32	53%	341	1.72	65%	510
CM (DiffLinker)	1.81	67%	347	1.59	61%	386	1.55	65%	508

Can We Generate Counterfactuals?

Although originally designed for masking tasks, CM can effectively generate realistic, chemically valid counterfactual examples by ensuring that molecules stay within the data distribution. In comparison to other counterfactual generation methods, it also enables targeted, local modifications to specific fragments, which can be easily interpreted by people with a chemistry background.



Dataset	Method	BSR ↑	Success Rate ↑	Similarity ↑	Diversity ↑	SA ↓
CYP 3A4	GNNExplainer	0.16 ± 0.04	0.16 ± 0.04	0.33 ± 0.00	0.00 ± 0.00	3.67 ± 0.04
	NN (Nearest Neighbour)	1.00 ± 0.00	1.00 ± 0.00	0.37 ± 0.00	0.66 ± 0.00	2.53 ± 0.01
	MMACE	0.27 ± 0.00	$\mathbf{1.00 \pm 0.00}$	$\mathbf{0.60 \pm 0.01}$	$\mathbf{0.52 \pm 0.01}$	3.34 ± 0.02
	CM (CReM)	0.21 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.52 ± 0.04	0.46 ± 0.04	$\mathbf{2.71 \pm 0.08}$
	CM (DiffLinker)	$\mathbf{0.41 \pm 0.02}$	$\mathbf{1.00 \pm 0.00}$	0.41 ± 0.05	0.41 ± 0.01	3.91 ± 0.18

arXiv

GitHub

Acknowledgments

This study was funded by the "Interpretable and Interactive Multimodal Retrieval in Drug Discovery" project. The "Interpretable and Interactive Multimodal Retrieval in Drug Discovery" project (FENG.02.02-IP.05-0040/23) is carried out within the First Team programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG). We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018272.