REALFLUX: A COUNTERFACTUAL EXPLANATION FRAMEWORK FOR MOLECULAR GRAPHS INSPIRED BY LOCAL OPTIMIZATION IN MEDICINAL CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

To address out-of-distribution (OOD) bias and topological leakage in graph masking, together with the decoupling of localization and generation in the evaluated Counterfactual Masking (CM) pipeline, we propose **RealFlux**, a unified molecular graph counterfactual explanation framework inspired by local optimization in medicinal chemistry. RealFlux introduces native directional utility (Native-U) for forward-utility-based localization without an external gradient explainer and combines Context-Locked Enforcement (CLE) with In-Distribution Gating (IDG) for constrained local intervention and predictor-representation validation. Across three pharmacological classification tasks, RealFlux achieves 97.7%–98.3% chemical validity and significantly higher classification flip rate than the evaluated CM baselines in all six pairwise comparisons after Holm correction (BSR, 0.286–0.680). On the common-substructure masking benchmark, the combined prediction discrepancy reaches $|\Delta\hat{y}| = 0.583$, with the same relative advantage across GIN, GraphSAGE, and GAT. Median per-molecule inference latency is 0.86 s, more than an order of magnitude lower than that of the evaluated 3D diffusion baseline. Overall, RealFlux provides a directionally controllable and computationally efficient framework for model-guided molecular counterfactual analysis.

1 INTRODUCTION

Graph neural networks (GNNs) can directly learn molecular representations from atom–bond graphs and have become important modeling tools for molecular property prediction and computational drug discovery Wieder et al. (2020). However, when such models are used for lead optimization, merely identifying "which atoms or substructures are important" is often insufficient to support downstream design decisions. An explanation must also address a question more closely aligned with medicinal chemistry practice: while preserving the molecular scaffold and data distribution as much as possible, where and how should the molecule be modified to shift the model prediction in a desired direction?

Existing post hoc explanation methods primarily rely on gradient attribution or subgraph-importance optimization, including Integrated Gradients, Grad-CAM, GNNExplainer, and SubgraphX Sundararajan et al. (2017); Selvaraju et al. (2017); Ying et al. (2019); Yuan et al. (2021); Sánchez-Lengeling et al. (2020). These methods are suitable for characterizing predictive sensitivity, but statistical importance does not necessarily imply chemical editability. Directly zeroing features, deleting nodes, or breaking chemical bonds in discrete molecular graphs may create graph states outside the training distribution. Moreover, structural information from the supposedly masked region may still leak through message passing if the original adjacency relationships remain accessible Janisiów et al. (2026). Counterfactual Masking (CM) alleviates this problem by replacing masked regions with generated fragments Janisiów et al. (2026). In the evaluated CM pipeline, an external explainer first localizes the region, CReM or DiffLinker then generates candidates, and the target model is used for downstream counterfactual selection Polishchuk (2020); Igashov et al. (2024); Janisiów et al. (2026). Localization and native candidate generation are therefore not driven by one shared target-direction utility in this comparison pipeline.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

RealFlux is motivated by local optimization in medicinal chemistry, where substitution, deletion, functional-group addition, or local scaffold adjustment is performed around an existing molecular core while preserving key structural elements as much as possible and iteratively moving toward a predefined property objective Zhao (2007); Lazzara & Moore (2020); Griffen et al. (2011). Following this principle, we reformulate the explanation problem as constrained directional optimization over an executable local chemical action space and introduce three complementary components. Native-U measures directional response within the feasible edit space; CLE treats the non-target contextual topology as a hard invariant; and IDG rejects candidates that are too distant from the reference training set in the target predictor's representation space. Together, these components place site localization, local intervention, and representation-space validation under a shared intervention semantics.

The main contributions of this work are summarized as follows:

1. **Action-space-aligned directional localization.** We introduce Native-U, which localizes candidate intervention regions by measuring the target-direction response of feasible first-step chemical edits. Localization is thus evaluated in the same executable edit space and on the same utility scale used for subsequent counterfactual intervention, without relying on an external gradient explainer.
2. **Topological and representation-space constraints for local counterfactuals.** CLE and IDG jointly constrain non-target topology and predictor-representation proximity. Relative to Feature Zeroing, RealFlux reduces combined-scenario $|\Delta\hat{y}|$ from 2.914 to 0.583; in the single-anchor setting, W decreases from 46.838 to 2.333.
3. **Generation-time directional control toward target properties.** Across three ADMET classification tasks, RealFlux achieves BSR values of 0.286–0.680. When each method's native generation output is frozen and no additional post-generation property-based reranking or reselection is applied, RealFlux obtains DCR values of 0.890–0.990.

2 Related Work

2.1 Molecular Graph Attribution and Masking

Molecular graph explanation methods commonly attribute model outputs to atoms, bonds, or subgraphs. Gradient-based approaches such as Integrated Gradients and Grad-CAM quantify local sensitivity in continuous representations, whereas GNNExplainer and SubgraphX search for structural subgraphs that preserve model predictions Sundararajan et al. (2017); Selvaraju et al. (2017); Ying et al. (2019); Yuan et al. (2021). Systematic evaluations of graph attribution methods have further shown that explanation quality should be examined along multiple dimensions, including accuracy, stability, fidelity, and consistency Sánchez-Lengeling et al. (2020). For molecular graphs, however, explanation operations themselves must also satisfy basic chemical constraints. Simple feature zeroing or structural deletion may alter the input distribution, while retained adjacency topology can still carry information from regions that are nominally "masked" Janisiów et al. (2026).

2.2 Molecular Counterfactual Explanations

Unlike attribution methods that merely identify important substructures, molecular counterfactual explanations search for minimal chemical edits that invert model predictions to guide lead optimization. Early approaches such as MMACE perturb SELFIES strings to generate local analogues, but do not impose the explicit hard context-locking invariant used by RealFlux Wellawatte et al. (2022). To mitigate the out-of-distribution bias of feature-zeroing masks, Counterfactual Masking (CM) pairs substructure removal with conditional fragment generation Janisiów et al. (2026). Specifically, CM deploys CReM, which mutates fragments within matching database contexts but restricts coverage by treating rings as indivisible units Polishchuk (2020); Janisiów et al. (2026), alongside DiffLinker, an E(3)-equivariant 3D diffusion model subject to point-cloud conversion and validity–speed trade-offs Igashov et al. (2024). In the evaluated CM pipeline, these generators operate separately from the target predictor; target-direction preference is therefore introduced at downstream selection rather than within their native sampling processes Polishchuk (2020); Igashov et al. (2024); Janisiów et al. (2026).

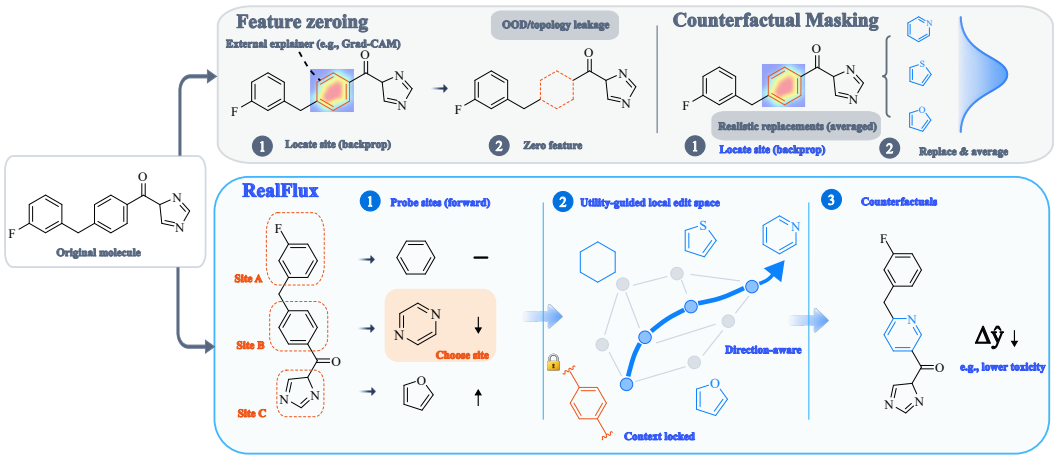

Figure 1: Paradigm comparison between RealFlux and the evaluated attribution–generation pipeline. The comparison pipeline first localizes a region using an external explainer, then employs a generator decoupled from the target model to produce candidates for post hoc filtering. RealFlux instead uses the same forward directional utility for site probing, constrained editing, and property validation, while locking the non-intervened context throughout the process.

2.3 Medicinal Chemistry Lead Optimization and Constrained Molecular Generation

In medicinal chemistry lead optimization, preservation of the molecular scaffold and analysis of local structure–activity relationships (SARs) are central principles Zhao (2007); Lazzara & Moore (2020); Griffen et al. (2011). Existing constrained generative models, including fragment-linking methods such as DeLinker and DiffLinker Imrie et al. (2020); Igashov et al. (2024) and rule-based approaches such as CReM Polishchuk (2020), support conditional generation around local attachment points. In the evaluated attribution–generation cascade, however, native candidate generation remains separate from the target predictor's directional utility. This motivates a narrower question: how can forward target utility be used consistently for local intervention while strictly preserving the non-target scaffold topology, without retraining a task-specific generative model?

3 Method

3.1 Problem Formulation and Framework

Let the input molecule be an attributed undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, F_v, F_e)$, and let the target predictor be $f : \mathcal{G} \to \mathbb{R}$. For binary classification, $f(\mathcal{G})$ denotes the probability of the target class; for regression, $f(\mathcal{G})$ denotes the predicted continuous property. The target direction is denoted by $\tau \in \{+1, -1\}$, where +1 indicates that the output should increase and -1 indicates that it should decrease.

Using BRICS decomposition together with common functional-group rules Degen et al. (2008), the molecule is decomposed into a set of candidate fragments $\mathcal{R} = \{r_1, \ldots, r_m\}$. For each candidate region r, we define a local chemical action space $\mathcal{A}(r)$ comprising executable replacement, deletion, and addition operations under the current attachment context and valence constraints. Any action $a \in \mathcal{A}(r)$ produces a locally modified molecule $\mathcal{G}' = a(\mathcal{G})$. RealFlux probes first-step feasible edits for all candidate regions and uses Native-U to identify the most promising region for intervention. It then continues local refinement along the same directional utility objective, considering only candidates jointly accepted by CLE and IDG, and ultimately returns the counterfactual molecule.

3.2 CORE COMPONENTS

3.2.1 NATIVE DIRECTIONAL UTILITY (NATIVE-U)

For a candidate $c = a(\mathcal{G})$, the target-direction log-odds utility for classification is defined as

$$u(c) = \tau \left[\operatorname{logit} f(c) - \operatorname{logit} f(\mathcal{G}_0)\right]. \tag{1}$$

For regression, the utility is normalized by the target standard deviation $\hat{\sigma}_y$ in the training set:

$$u(c) = \frac{\tau[f(c) - f(\mathcal{G}_0)]}{\hat{\sigma}_y}. \tag{2}$$

Thus, $u(c) > 0$ indicates that the edit shifts the prediction toward the desired direction. For a region r, we enumerate its first-step feasible candidates $\mathcal{C}_{\text{first}}(r)$ and define

$$\text{score}(r) = -\frac{1}{|\mathcal{C}_{\text{first}}(r)|} \sum_{c \in \mathcal{C}_{\text{first}}(r)} u(c). \tag{3}$$

Because of the leading minus sign, lower scores correspond to larger mean positive directional utility; RealFlux therefore selects $r^* = \arg\min_{r \in \mathcal{R}} \text{score}(r)$.

3.2.2 CONTEXT-LOCKED ENFORCEMENT (CLE)

For an editable fragment r, its complementary context is defined as $\mathcal{C}(\mathcal{G}, r) = \mathcal{G} \setminus r$, and the set of boundary anchors is

$$A = \{v \in \mathcal{C}(\mathcal{G}, r) : \exists u \in r,\, (v, u) \in \mathcal{E}\}. \tag{4}$$

CLE permits only local actions satisfying the following invariants:

$$\mathcal{A}_{\text{CLE}}(r) = \{a \in \mathcal{A}(r) \mid \mathcal{G}'[\mathcal{V}_{\mathcal{C}}] \cong \mathcal{C}(\mathcal{G}, r),\ \text{Arom}(\mathcal{G}'[\mathcal{V}_{\mathcal{C}}]) = \text{Arom}(\mathcal{C}(\mathcal{G}, r))\}. \tag{5}$$

Except for the replaced region and its boundary bonds to the anchors, the atom types, connectivity topology, and aromaticity states of the non-intervened region remain unchanged. CLE therefore restricts counterfactual differences to the explicitly declared intervention region, ensuring that observed prediction changes correspond to constrained local structural modifications.

3.2.3 IN-DISTRIBUTION GATING (IDG) AND CHEMICAL FEASIBILITY

Let $h(\mathcal{G}) \in \mathbb{R}^D$ denote the internal graph representation of the target predictor, and let the reference training embeddings be

$$\mathcal{Z}_{\text{ref}} = \{h(\mathcal{G}_i)\}. \tag{6}$$

For a candidate $\mathcal{G}'$, the nearest-neighbor distance is defined as

$$D_{\text{ref}}(\mathcal{G}') = \min_{z \in \mathcal{Z}_{\text{ref}}} \|h(\mathcal{G}') - z\|_2. \tag{7}$$

The threshold is set to the 90th percentile of nearest-neighbor distances in the validation set:

$$\tau_{\text{IDG}} = Q_{0.90}\left(\{D_{\text{ref}}(\mathcal{G}_{\text{val}})\}\right). \tag{8}$$

A candidate is rejected when $D_{\text{ref}}(\mathcal{G}') > \tau_{\text{IDG}}$. Accordingly, IDG assumes access to this predictor representation in addition to forward predictions. All intermediate structures must additionally

Table 1: Comprehensive comparison of molecular counterfactual generation quality

Dataset	Method	Chemical Validity ↑	BSR ↑	Success Rate ↑	Similarity ↑	Diversity ↑	SA ↓
CYP3A4	CM (CReM)	0.600 ± 0.090	0.332 ± 0.013	0.395 ± 0.038	$\mathbf{0.540 \pm 0.019}$	$\mathbf{0.461 \pm 0.020}$	$\mathbf{2.680 \pm 0.049}$
	CM (DiffLinker)	0.886 ± 0.008	0.370 ± 0.044	0.412 ± 0.038	0.446 ± 0.058	0.291 ± 0.019	3.782 ± 0.184
	RealFlux	$\mathbf{0.979 \pm 0.003}$	$\mathbf{0.638 \pm 0.015}$	$\mathbf{0.843 \pm 0.019}$	0.473 ± 0.009	0.446 ± 0.018	3.395 ± 0.013
CYP2D6	CM (CReM)	0.639 ± 0.037	0.099 ± 0.023	0.241 ± 0.022	0.448 ± 0.014	$\mathbf{0.504 \pm 0.012}$	$\mathbf{2.656 \pm 0.039}$
	CM (DiffLinker)	0.861 ± 0.014	0.131 ± 0.015	0.146 ± 0.004	$\mathbf{0.450 \pm 0.012}$	0.318 ± 0.022	3.785 ± 0.160
	RealFlux	$\mathbf{0.977 \pm 0.001}$	$\mathbf{0.286 \pm 0.051}$	$\mathbf{0.498 \pm 0.080}$	0.425 ± 0.009	0.383 ± 0.016	3.529 ± 0.045
hERG	CM (CReM)	0.441 ± 0.038	0.256 ± 0.011	0.346 ± 0.017	0.447 ± 0.018	$\mathbf{0.522 \pm 0.015}$	$\mathbf{3.165 \pm 0.038}$
	CM (DiffLinker)	0.872 ± 0.006	0.475 ± 0.044	0.407 ± 0.033	0.434 ± 0.007	0.363 ± 0.015	4.070 ± 0.064
	RealFlux	$\mathbf{0.983 \pm 0.002}$	$\mathbf{0.680 \pm 0.026}$	$\mathbf{0.857 \pm 0.009}$	$\mathbf{0.504 \pm 0.003}$	0.416 ± 0.012	3.813 ± 0.031

Note: Values are pool-level means $\pm$ standard deviations across three independent random seeds (Seeds 5, 15, and 25; $\mathrm{ddof} = 0$). Bold indicates the best result in the corresponding column. BSR is computed as a pooled ratio over the full candidate pool. Pairwise comparisons are corrected for multiple testing using the Holm–Bonferroni procedure Holm (1979).

pass RDKit sanitization and valence checks, consist of a single connected component, and satisfy $\mathrm{SA}(\mathcal{G}') \leq 4.5$ Ertl & Schuffenhauer (2009). Among candidates satisfying CLE, IDG, and these chemical-feasibility constraints, RealFlux ranks edits using the same $u(c)$ used during localization, so localization and edit evaluation share one directional objective.

4 Experimental Setup

We evaluate RealFlux on the CYP3A4, CYP2D6, and hERG classification tasks associated with TDC, together with common-substructure pairs constructed from the AqSolDB aqueous-solubility dataset Huang et al. (2021; 2022); Veith et al. (2009); Karim et al. (2021); Sorkun et al. (2019). The primary predictor is a three-layer GIN with a hidden dimension of 512, global mean pooling, and dropout of 0.3 Xu et al. (2019). Cross-architecture validation additionally uses GraphSAGE and GAT Hamilton et al. (2017); Veličković et al. (2018). Classification datasets are split 80/10/10 using Seeds 5, 15, and 25, with 500 inputs evaluated per dataset per seed. The masking benchmark contains 881 PubChem superstructure–molecule pairs from AqSolDB, including a single-anchor subset of 516 pairs. Baselines include Feature Zeroing, CM(CReM), and CM(DiffLinker). All methods use a 300 s per-input time limit; method-specific execution procedures are described in Appendix A.

The core metrics include Chemical Validity; Before Success Rate (BSR, $\sum \mathrm{flip} / \sum \mathrm{valid}$); Success Rate, defined as the proportion of inputs for which at least one valid flipped molecule is obtained; Morgan-fingerprint Tanimoto Similarity and Diversity Rogers & Hahn (2010); Bajusz et al. (2015); SA Ertl & Schuffenhauer (2009); Directional Margin Score (DMS) and Direction Consistency Rate (DCR) under a frozen native-generation-pool protocol without additional post-generation property-based reranking or reselection; and prediction leakage $|\Delta\hat{y}|$, together with latent-space Wasserstein distance W, on common-substructure pairs. Pairwise differences in BSR are assessed using 10,000 sample-level paired bootstrap replicates with Holm–Bonferroni correction Holm (1979). DCR is reported with Wilson 95% confidence intervals Wilson (1927). Full data-processing procedures, metric definitions, and baseline execution protocols are provided in Appendix A.

5 Results and Analysis

5.1 Counterfactual Quality and Directional Control

Table 1 evaluates counterfactual generation performance on the CYP3A4, CYP2D6, and hERG datasets, with 500 test molecules per dataset. Results are reported as the mean $\pm$ standard deviation across three independent random seeds.

RealFlux achieves BSR values of 0.638, 0.286, and 0.680 on the three datasets, respectively. The absolute improvements over CM(CReM) are +0.306, +0.187, and +0.424, while those over CM(DiffLinker) are +0.268, +0.155, and +0.205. All six pairwise comparisons remain significant after Holm correction (adjusted $p < 0.001$). Meanwhile, the chemical validity of RealFlux remains stable at 0.977–0.983. These gains do not imply dominance on every metric: CReM has lower SA and higher diversity on all three tasks, and on CYP3A4 it also has higher similarity (0.540 versus

Table 2: Directional controllability with no additional post-generation property-based reranking or reselection (100-molecule test set)

Dataset	Method	DMS Mean [95% CI] ↑	DCR [95% CI] ↑
CYP3A4	CM (CReM)	+0.0307 [0.0122, 0.0496]	0.650 [0.553, 0.736]
	CM (DiffLinker)	+0.1317 [0.0991, 0.1653]	0.740 [0.646, 0.816]
	RealFlux	**+0.3933 [+0.3527, +0.4320]**	**0.890 [0.814, 0.937]**
CYP2D6	CM (CReM)	+0.0325 [0.0160, 0.0487]	0.600 [0.502, 0.691]
	CM (DiffLinker)	$-0.0041[-0.0391, +0.0334]$	0.250 [0.176, 0.343]
	RealFlux	**+0.2139 [+0.1805, +0.2482]**	**0.940 [0.875, 0.972]**
hERG	CM (CReM)	+0.0881 [0.0628, 0.1143]	0.670 [0.573, 0.754]
	CM (DiffLinker)	+0.2816 [0.1935, 0.3708]	0.530 [0.433, 0.625]
	RealFlux	**+0.4923 [+0.4532, +0.5302]**	**0.990 [0.946, 0.998]**

Note: For each method, the native generation output is frozen before evaluation-stage property scoring; no additional post-generation property-based reranking or reselection is applied. The 95% confidence intervals for DMS are based on 10,000 bootstrap replicates; those for DCR are Wilson intervals Wilson (1927).

Table 3: Evaluation of paired common-substructure masking and topological information leakage on AqSolDB

Metric	No masking	Feature zeroing	CM (CReM)	CM (DiffLinker)	RealFlux
Single Anchor $\lvert\Delta\hat{y}\rvert \downarrow$	2.694	2.919	1.735	1.184	**0.253**
Single Anchor Validity ↑	1.000	1.000	0.886	0.649	0.975
Single Anchor Pairs	516	516	457	335	503
Single Anchor $W \downarrow$	**2.245**	46.838	2.414	6.620	2.333
Multiple Anchors $\lvert\Delta\hat{y}\rvert \downarrow$	2.421	2.147	1.519	1.101	**0.168**
Multiple Anchors Validity ↑	1.000	1.000	0.538	0.591	0.900
Multiple Anchors Pairs	667	667	359	394	600
Combined $\lvert\Delta\hat{y}\rvert \downarrow$	4.135	2.914	2.480	1.880	**0.583**
Combined Validity ↑	1.000	1.000	0.566	0.604	0.828
Combined Pairs	873	873	494	527	723

Note: The Wasserstein distance W is computed from embeddings extracted from the penultimate GNN layer. Equal-size random downsampling is performed using the shared minimum valid sample size of $n_{\min} = 1{,}032$ across methods.

0.473). The results therefore indicate a trade-off between stronger property-flipping performance, local structural conservatism, diversity, and empirically estimated synthetic difficulty.

To assess whether directional bias persists without additional post-generation property-based selection, we evaluate a frozen native-generation protocol ($N = 100$): each method's native generation output is fixed before evaluation-stage scoring, and no additional property-based reranking or reselection is applied (Table 2).

Without additional post-generation property-based reranking or reselection, the evaluated CM baselines show weaker or less consistent directional control under this protocol. For example, on CYP2D6, CM(DiffLinker) obtains a DCR of 0.250 and a negative mean DMS. RealFlux obtains DCR values of 0.890–0.990 across the three tasks; all corresponding 95% confidence-interval lower bounds exceed 0.5, and all DMS intervals remain above zero. These results support the narrower claim that directional bias is already present in RealFlux's native generation output and does not arise solely from an additional post-generation property-based selection step.

5.2 Masking Fidelity and Cross-Architecture Robustness

Using common-substructure pairs constructed from the AqSolDB aqueous-solubility dataset Sorkun et al. (2019), we systematically evaluate the fidelity of graph-masking strategies. An ideal masking approach should completely obscure structural features from the non-shared region, causing the absolute difference in paired predictions, $|\Delta\hat{y}|$, to approach zero; this paired common-substructure evaluation follows the masking-fidelity logic introduced for molecular CM Janisiów et al. (2026). At the same time, generated molecules should remain close to the reference distribution in the predictor's latent representation space, as reflected by a small Wasserstein distance W.

As shown in Table 3, conventional Feature Zeroing retains graph connections in the non-shared region and produces a prediction discrepancy of $|\Delta\hat{y}|$=2.914 in the combined setting, indicating that

Table 4: Ablation study of the core RealFlux components.

Metric	**RealFlux (Full)**	w/o CLE	w/o IDG	w/o Native-U
Context Retention ↑	0.654 ± 0.064	0.233 ± 0.076	0.664 ± 0.060	0.658 ± 0.057
Targeted Success (TS) ↑	0.622 ± 0.070	0.250 ± 0.076	0.644 ± 0.064	0.478 ± 0.113
Prediction Leakage $\lvert\Delta\hat{y}\rvert$ ↓	0.256 ± 0.013	0.184 ± 0.011	0.197 ± 0.008	0.238 ± 0.014
Distribution Distance (W) ↓	2.529 ± 0.162	3.125 ± 0.238	4.812 ± 0.458	2.148 ± 0.186
Consistency ↑	0.748 ± 0.068	0.727 ± 0.065	0.748 ± 0.068	0.672 ± 0.049
Chemical Validity ↑	0.979 ± 0.003	0.979 ± 0.003	0.991 ± 0.005	0.977 ± 0.004
Latency P50 ↓ (s)	0.862 ± 0.005	0.950 ± 0.184	0.813 ± 0.139	2.557 ± 0.069

zeroing node features alone does not sufficiently remove the influence of the non-shared region on the model output; preservation of adjacency under feature masking is a known source of topological leakage in molecular graph explanations Janisiów et al. (2026). In contrast, RealFlux reduces the prediction discrepancy to 0.253 and 0.168 in the single-anchor and multiple-anchor settings, respectively, and to 0.583 in the combined setting. This corresponds to an 80.0% reduction relative to Feature Zeroing and a 76.5% reduction relative to CM(CReM). Paired statistical tests further show that RealFlux provides significantly stronger control of prediction discrepancy than all comparison baselines ($p < 0.0001$).

In terms of predictor-representation proximity, Feature Zeroing creates nonphysical zero-feature subgraphs and increases the latent-space distribution distance to $W = 46.838$, consistent with the OOD concern documented for zero-feature molecular masking Janisiów et al. (2026). CM(DiffLinker) yields $W = 6.620$, whereas RealFlux achieves $W = 2.333$, close to the reference value of 2.245 for unmasked molecules. Because IDG and W both use predictor-derived representations, W is interpreted here as an internal representation-space diagnostic rather than an independent OOD test. Complete results across GIN, GraphSAGE, and GAT are provided in Appendix B. RealFlux achieves the lowest combined $\lvert\Delta\hat{y}\rvert$ under all three architectures, indicating that this behavior is not specific to GIN.

5.3 Mechanistic Ablation of Core Components

To distinguish the contributions of individual design components to overall performance, Table 4 summarizes the principal single-factor ablation results obtained across three random seeds (Seeds 5, 15, and 25).

The ablations support distinct roles for the three components, but also show explicit trade-offs. Removing CLE reduces Context Retention from 0.654 ± 0.064 to 0.233 ± 0.076 and Targeted Success from 0.622 to 0.250, consistent with CLE's role in preserving the non-intervened context. Removing IDG increases W from 2.529 ± 0.162 to 4.812 ± 0.458, while Targeted Success increases slightly and prediction leakage decreases; IDG therefore improves predictor-representation proximity rather than uniformly improving every task metric. Replacing Native-U with Grad-CAM reduces Targeted Success from 0.622 to 0.478 and Consistency from 0.748 to 0.672, while increasing P50 latency from 0.862 s to 2.557 s; the corresponding W is lower in this ablation. Together with the marginal-contribution rank-correlation results in Appendix C, these findings are consistent with CLE, IDG, and Native-U serving context preservation, representation-space gating, and actionability-aware directional localization, respectively.

5.4 Case Study

Three molecules with experimentally recorded activity were selected from public CYP450 inhibition data (PubChem CIDs 667700, 3233055, and 3240659) Veith et al. (2009); Huang et al. (2021). In all cases, the optimization objective was to reduce the model-predicted risk of CYP3A4 inhibition while preserving a specified core ring system. These examples are qualitative illustrations of model-predicted counterfactual changes; the generated molecules themselves are not experimentally validated in this study.

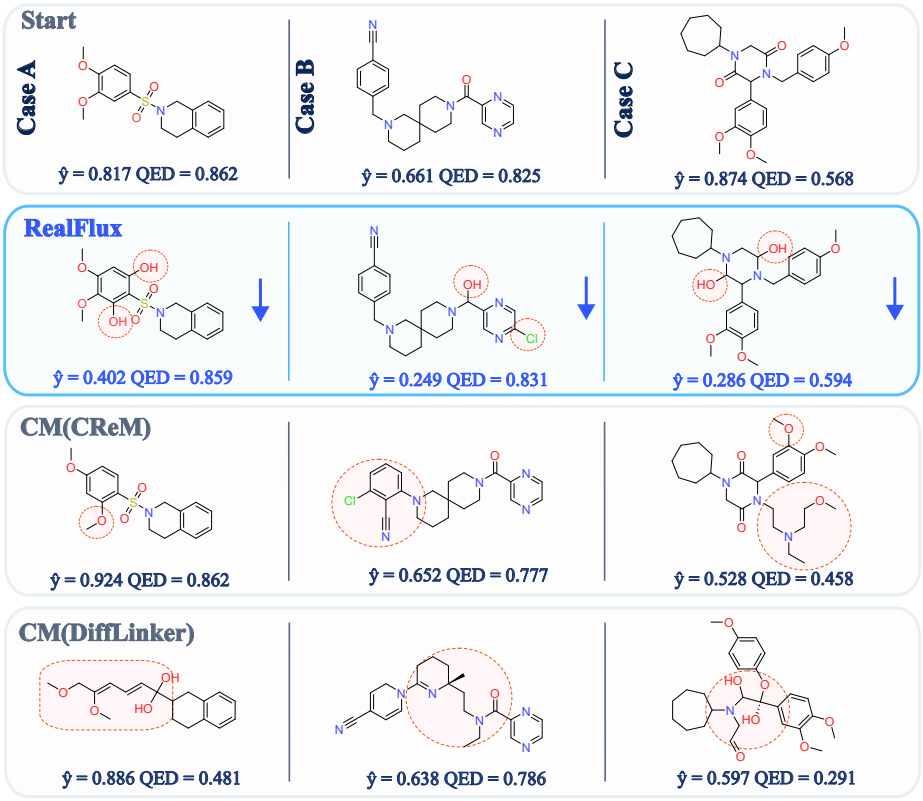

Figure 2: Comparison of local structural optimization for three CYP3A4 benchmark cases. The figure shows the molecular structures produced by the different methods for three representative scaffolds (Cases A–C), together with the predicted inhibition probability ($\hat{y}$) and quantitative estimate of drug-likeness (QED) Bickerton et al. (2012). Red dashed boxes indicate local editing sites, and blue arrows indicate predicted inhibition probabilities crossing the model decision threshold ($\hat{y} \leq 0.5$).

Case A. The starting molecule contains a 3,4-dimethoxybenzenesulfonyl group. Related 3,4-dimethoxyphenyl systems have documented oxidative metabolic liabilities, including CYP-associated O-demethylation, while medicinal-chemistry studies have also highlighted reactive-metabolite considerations for this substitution pattern Nelson et al. (1988); Crawford et al. (2012). While preserving the scaffold, RealFlux introduces two phenolic hydroxyl groups, reducing $\hat{y}$ from 0.817 to 0.402 while changing QED only from 0.862 to 0.859. In contrast, CM(CReM) only modifies the methoxy substitution site and increases $\hat{y}$ to 0.924, whereas the candidate generated by CM(DiffLinker) alters the benzene-ring scaffold ($\hat{y}$=0.886, QED=0.481).

Case B. Pyrazine-containing motifs can undergo oxidative metabolism, including ring hydroxylation, in microsomal or in vivo systems Hawksworth & Scheline (1975); Edlund & Baranczewski (2004). RealFlux converts the carbonyl group to a hydroxymethyl group and introduces chlorination on the pyrazine ring, reducing $\hat{y}$ to 0.249 while maintaining QED at 0.831. The candidate produced by CM(CReM) is modified primarily at the distal benzonitrile site, and $\hat{y}$ remains in the high-inhibition range at 0.652. CM(DiffLinker), by contrast, produces a ring-opened structure derived from the spirocyclic scaffold ($\hat{y}$=0.638, QED=0.786).

Case C. The 2,5-diketopiperazine scaffold is conformationally defined and comparatively rigid, and the accessibility of an imidazole nitrogen can strongly influence coordination to cytochrome P450 heme iron Ressurreição et al. (2011); Perrin et al. (2009); Aninat et al. (2001). RealFlux reduces both carbonyl groups, decreasing $\hat{y}$ from 0.874 to 0.286, corresponding to a 67.3% reduction, while increasing QED from 0.568 to 0.594. The best candidate generated by CM(CReM) reduces $\hat{y}$ only to 0.528, with modifications confined to the peripheral alkyl side chain. CM(DiffLinker), in contrast, causes ring-opening cleavage of the diketopiperazine scaffold ($\hat{y}$=0.597, QED=0.291).

6 LIMITATIONS

RealFlux has several current limitations and boundaries of applicability.

Empirical Synthetic-Accessibility Assessment. Although product-level synthetic accessibility thresholds are incorporated through the SA constraint Ertl & Schuffenhauer (2009), the generated candidates still exhibit higher average estimated synthetic difficulty than candidates produced through database-based fragment recombination. For individual molecules lacking an evident synthetic route, their practical synthesizability still requires expert assessment and experimental validation.

Evaluation Scope. Directional counterfactual generation is evaluated on three classification tasks under the reported 80/10/10 random splits. Cross-architecture analyses cover GIN, GraphSAGE, and GAT, but the study does not establish the same conclusions for alternative split protocols, directional regression generation, SMILES-based molecular language models, or all-atom equivariant predictors.

Limited 3D Ring Reconstruction. Complex partial-ring edits involving bridged, spirocyclic, or fused polycyclic systems remain challenging for the discrete macro-editing strategy. The current action space does not explicitly model ring strain or three-dimensional steric relaxation, so such restructuring may benefit from continuous three-dimensional geometric representations, as exemplified by DiffLinker Igashov et al. (2024). The complete four-level stress test is reported in Appendix D.

7 CONCLUSION

RealFlux unifies structural localization, constrained intervention, and property validation for molecular graph counterfactual explanations within a discrete local chemical action space. Native-U aligns actionability-aware site localization and edit evaluation under the same target-direction objective, while CLE and IDG constrain non-target topology and predictor-representation deviation. Across the evaluated tasks and architectures, RealFlux improves classification-flip performance, reduces post-masking prediction discrepancies, and provides lower measured latency than the evaluated 3D diffusion baseline, while retaining the reported trade-offs in similarity, diversity, and synthetic accessibility. When native generation outputs are frozen, the directional bias persists without additional post-generation property-based reranking or reselection. These results support RealFlux as a unified constrained-search framework for model-guided molecular counterfactual analysis within the evaluated setting.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing as well as for retrieval and discovery (e.g., finding related work). We have not used generative AI tools for core methodology design, algorithm implementation, or experimental execution. All AI-assisted text revisions, literature references, and retrieved details were manually reviewed and verified by the authors. The authors take full responsibility for the final content, claims, and artifacts presented in this paper.

ETHICS STATEMENT

This study focuses on computational interpretability for molecular property prediction and uses publicly available chemical datasets (Therapeutics Data Commons and AqSolDB). It does not involve human subjects, vertebrate animal experiments, or proprietary clinical data. The generated candidates are evaluated computationally and are not experimentally synthesized or validated in this study.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 specifies the method definitions and constraints; Section 4 and Appendix A describe the datasets, splits, metrics, baseline procedures, statistical tests, and random

seeds (5, 15, and 25). Appendices B–D report the cross-architecture, localization, and topology-stress evaluations.

REFERENCES

Caroline Aninat, Yoshio Hayashi, François André, and Marcel Delaforge. Molecular requirements for inhibition of cytochrome P450 activities by roquefortine. *Chemical Research in Toxicology*, 14(9):1259–1265, 2001. doi: 10.1021/tx015512l. URL https://doi.org/10.1021/tx0155121.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:20, 2015. doi: 10.1186/s13321-015-0069-3. URL https://doi.org/10.1186/s13321-015-0069-3.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012. doi: 10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

James J. Crawford, Peter W. Kenny, Jonathan Bowyer, Calum R. Cook, Jonathan E. Finlayson, Christine Heyes, Adrian J. Highton, Julian A. Hudson, Anja Jestel, Stephan Krapp, Scott Martin, Philip A. Macfaul, Benjamin P. McDermott, Thomas M. McGuire, Andrew D. Morley, Jeffrey J. Morris, Ken M. Page, Lyn Rosenbrier Ribeiro, Helen Sawney, Stefan Steinbacher, Caroline Smith, and Alexander G. Dossetter. Pharmacokinetic benefits of 3,4-dimethoxy substitution of a phenyl ring and design of isosteres yielding orally available cathepsin k inhibitors. *Journal of Medicinal Chemistry*, 55(20):8827–8837, 2012. doi: 10.1021/jm301119s. URL https://doi.org/10.1021/jm301119s.

Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008. doi: 10.1002/cmdc.200800178. URL https://doi.org/10.1002/cmdc.200800178.

Per Olof Edlund and Paweł Baranczewski. Identification of BVT.2938 metabolites by LC/MS and LC/MS/MS after in vitro incubations with liver microsomes and hepatocytes. *Journal of Pharmaceutical and Biomedical Analysis*, 34(5):1079–1090, 2004. doi: 10.1016/j.jpba.2003.12.010. URL https://doi.org/10.1016/j.jpba.2003.12.010.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1:8, 2009. doi: 10.1186/1758-2946-1-8. URL https://doi.org/10.1186/1758-2946-1-8.

Ed Griffen, Andrew G. Leach, Graeme R. Robb, and Daniel J. Warner. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry*, 54(22):7739–7750, 2011. doi: 10.1021/jm200452d. URL https://doi.org/10.1021/jm200452d.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1024–1034, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Gabrielle Hawksworth and Ronald R. Scheline. Metabolism in the rat of some pyrazine derivatives having flavour importance in foods. *Xenobiotica*, 5(7):389–399, 1975. doi: 10.3109/00498257509056108. URL https://doi.org/10.3109/00498257509056108.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/4c56ff4ce4aaf9573aa5dff913df997a-Abstract-round1.html.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 18:1033–1036, 2022. doi: 10.1038/s41589-022-01131-2. URL https://doi.org/10.1038/s41589-022-01131-2.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3D-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, 6: 417–427, 2024. doi: 10.1038/s42256-024-00815-9. URL https://doi.org/10.1038/s42256-024-00815-9.

Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4): 1983–1995, 2020. doi: 10.1021/acs.jcim.9b01120. URL https://doi.org/10.1021/acs.jcim.9b01120.

Łukasz Janisiów, Marek Kochańczyk, Bartosz Michał Zieliński, and Tomasz Danel. Enhancing chemical explainability through counterfactual masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22155–22163, 2026. doi: 10.1609/aaai.v40i26.39371. URL https://doi.org/10.1609/aaai.v40i26.39371.

Abdul Karim, Matthew Lee, Thomas Balle, and Abdul Sattar. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *Journal of Cheminformatics*, 13:60, 2021. doi: 10.1186/s13321-021-00541-z. URL https://doi.org/10.1186/s13321-021-00541-z.

Phillip R. Lazzara and Terry W. Moore. Scaffold-hopping as a strategy to address metabolic liabilities of aromatic compounds. *RSC Medicinal Chemistry*, 11:18–29, 2020. doi: 10.1039/C9MD00396G. URL https://doi.org/10.1039/C9MD00396G.

W. L. Nelson, L. D. Olsen, D. B. Beitner, and R. J. Jr. Pallow. Regiochemistry and substrate stereoselectivity of o-demethylation of verapamil in the presence of the microsomal fraction from rat and human liver. *Drug Metabolism and Disposition*, 16(2):184–188, 1988. URL https://pubmed.ncbi.nlm.nih.gov/2898330/.

Lionel Perrin, François André, Caroline Aninat, Rémy Ricoux, Jean-Pierre Mahy, Ning Shangguan, Madeleine M. Joullié, and Marcel Delaforge. Intramolecular hydrogen bonding as a determinant of the inhibitory potency of n-unsubstituted imidazole derivatives towards mammalian hemoproteins. *Metallomics*, 1(2):148–156, 2009. doi: 10.1039/B817743K. URL https://doi.org/10.1039/B817743K.

Pavel Polishchuk. CReM: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, 12:28, 2020. doi: 10.1186/s13321-020-00431-w. URL https://doi.org/10.1186/s13321-020-00431-w.

Ana Sofia M. Ressurreição, Régis Delatouche, Cesare Gennari, and Umberto Piarulli. Bifunctional 2,5-diketopiperazines as rigid three-dimensional scaffolds in receptors and peptidomimetics. *European Journal of Organic Chemistry*, 2011(2):217–228, 2011. doi: 10.1002/ejoc.201001330. URL https://doi.org/10.1002/ejoc.201001330.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Benjamín Sánchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, and Alexander B. Wiltschko. Evaluating attribution for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/417fbbf2e9d5a28a855a11894b2e795a-Abstract.html.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6:143, 2019. doi: 10.1038/s41597-019-0151-1. URL https://doi.org/10.1038/s41597-019-0151-1.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Henrike Veith, Noel Southall, Ruili Huang, Tim James, Darren Fayne, Natalia Artemenko, Min Shen, James Inglese, Christopher P. Austin, David G. Lloyd, and Douglas S. Auld. Comprehensive characterization of cytochrome P450 isozyme selectivity across chemical libraries. *Nature Biotechnology*, 27(11):1050–1055, 2009. doi: 10.1038/nbt.1581. URL https://doi.org/10.1038/nbt.1581.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science*, 13:3697–3705, 2022. doi: 10.1039/D1SC05259D. URL https://doi.org/10.1039/D1SC05259D.

Oliver Wieder, Stefan Kohlbacher, Mélaine Kuenemann, Arthur Garon, Pierre Ducrot, Thomas Seidel, and Thierry Langer. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technologies*, 37:1–12, 2020. doi: 10.1016/j.ddtec.2020.11.009. URL https://doi.org/10.1016/j.ddtec.2020.11.009.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://doi.org/10.1080/01621459.1927.10502953.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9240–9251, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12241–12252. PMLR, 2021. URL https://proceedings.mlr.press/v139/yuan21c.html.

Hongyu Zhao. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discovery Today*, 12(3–4):149–155, 2007. doi: 10.1016/j.drudis.2006.12.003. URL https://doi.org/10.1016/j.drudis.2006.12.003.

A Experimental and Evaluation Details

A.1 Datasets and Predictive Models

The classification experiments use CYP3A4, CYP2D6, and hERG Huang et al. (2021; 2022); Veith et al. (2009); Karim et al. (2021). Each dataset is split into training, validation, and test partitions using an 80/10/10 ratio. Three fixed random seeds (5, 15, and 25) are used, with 500 input molecules evaluated for each seed. The aqueous-solubility masking experiments use AqSolDB Sorkun et al. (2019) to construct common-substructure pairs, comprising 881 PubChem superstructure–molecule

pairs, including 516 single-anchor pairs. The primary predictor is a three-layer GIN with hidden size 512, global mean pooling, and dropout 0.30 Xu et al. (2019). GraphSAGE and GAT are additionally used for cross-architecture verification Hamilton et al. (2017); Veličković et al. (2018).

A.2 Baselines and Execution Budget

CM(CReM) performs local replacement using fragments associated with co-occurring chemical environments Polishchuk (2020); Janisiów et al. (2026). CM(DiffLinker) performs fragment linking through three-dimensional equivariant diffusion Igashov et al. (2024); Janisiów et al. (2026). Conventional masking baselines include Feature Zeroing and No Masking. Baseline-specific generation procedures follow the corresponding comparison setup. All methods use the same time limit of 300 s per input.

A.3 Metrics and Statistics

Chemical Validity: the proportion of molecules that pass RDKit parsing, sanitization, valence checks, and filtering for a single connected component.

BSR: the proportion of target-class flips within the original valid candidate pool before downstream similarity- or property-based refiltering, computed as the pooled ratio $\sum \mathrm{flip} / \sum \mathrm{valid}$.

Success Rate: the proportion of inputs for which at least one valid flipped counterfactual is obtained after application of the unified selector.

Similarity / Diversity: Tanimoto similarity between counterfactuals and the original molecule, and internal diversity among counterfactuals, both computed using Morgan fingerprints and reported only for successfully flipped samples Rogers & Hahn (2010); Bajusz et al. (2015).

SA: synthetic accessibility score, where lower values indicate lower empirically estimated synthetic difficulty Ertl & Schuffenhauer (2009).

DMS / DCR: under the frozen native-generation-pool protocol with no additional post-generation property-based reranking or reselection, DMS measures the mean directional log-odds displacement, while DCR measures the proportion of prediction changes consistent with the specified direction.

$|\Delta\hat{y}|$**/W:** for common-substructure pairs, the former measures the absolute paired prediction difference after masking, whereas the latter measures the Wasserstein distance between candidate and reference distributions in the predictor-derived embedding space.

Pairwise BSR comparisons use 10,000 sample-level paired bootstrap replicates with Holm–Bonferroni correction for multiple comparisons Holm (1979). The 95% confidence interval for DMS is obtained by bootstrap, whereas DCR uses Wilson intervals Wilson (1927). Wasserstein distance is computed after equal-size random downsampling using the shared minimum valid sample size $n_{\min} = 1{,}032$.

B Cross-Architecture Masking Robustness

Table 3 in the main text reports the core masking results under GIN. To examine whether the conclusions depend on a particular message-passing architecture, complete results for GIN, GraphSAGE, and GAT Xu et al. (2019); Hamilton et al. (2017); Veličković et al. (2018) are provided below.

As shown in Table 5, across all three architectures, RealFlux remains significantly better on the combined $|\Delta\hat{y}|$ metric after Holm correction, with no reversal in relative performance. For distribution distance W, RealFlux remains significantly lower than CM(DiffLinker) under every architecture. These results indicate that the relative advantage of RealFlux in controlling prediction discrepancies is maintained across the tested message-passing architectures.

C Native-U vs. Grad-CAM

To evaluate the alignment of Native-U with observed local intervention effects without an external explainer, this experiment uses CReM-generated local interventions Polishchuk (2020) as the refer-

Table 5: Cross-model masking robustness across three graph neural network architectures

Architecture	Masking Method	Single Anchor $\lvert\Delta\hat{y}\rvert \downarrow$	Multi Anchor $\lvert\Delta\hat{y}\rvert \downarrow$	Combined $\lvert\Delta\hat{y}\rvert \downarrow$	Single Anchor $W \downarrow$
GIN	Feature zeroing	2.919	2.147	2.914	46.838
	CM (CReM)	1.735	1.519	2.480	2.414
	CM (DiffLinker)	1.184	1.101	1.880	6.620
	RealFlux	**0.253**	**0.168**	**0.583**	**2.333**
GraphSAGE	Feature zeroing	0.887	0.996	1.192	4.412
	CM (CReM)	1.681	1.461	2.425	**3.515**
	CM (DiffLinker)	1.140	0.889	1.514	8.113
	RealFlux	**0.223**	**0.145**	**0.524**	3.646
GAT	Feature zeroing	1.131	1.037	1.199	6.161
	CM (CReM)	1.744	1.490	2.488	**4.623**
	CM (DiffLinker)	1.185	0.875	1.440	8.953
	RealFlux	**0.230**	**0.157**	**0.557**	4.801

Table 6: Cross-architecture comparison of Native-U and external Grad-CAM attribution

Architecture	Attribution Method	Mean Consistency [Lower CI] ↑	Marginal Rank ρ [Lower CI] ↑
GIN	Grad-CAM	0.642	0.064
	RealFlux Native-U	**0.774 [0.737]**	**0.458 [0.308]**
GraphSAGE	Grad-CAM	0.551	0.130
	RealFlux Native-U	**0.596 [0.551]**	**0.480 [0.328]**
GAT	Grad-CAM	0.551	0.157
	RealFlux Native-U	**0.593 [0.547]**	**0.510 [0.361]**

ence and compares Native-U with Grad-CAM Selvaraju et al. (2017) in terms of prediction-direction consistency and marginal-contribution rank correlation (Marginal Rank ρ) (Table 6).

Under GIN, Native-U increases prediction-direction consistency from 0.642 to 0.774 and the marginal-rank correlation ρ from 0.064 to 0.458. Under GraphSAGE and GAT, Native-U also yields higher reported rank correlations, with $\rho = 0.480$ and $\rho = 0.510$, respectively. Within this intervention-based evaluation, Native-U is more closely aligned with the observed directional and marginal ordering of local edits than Grad-CAM.

D Topology, Efficiency, and Failure Boundaries

To systematically assess the topological adaptability and methodological boundaries of RealFlux, structural interventions are extended to four topological levels of increasing geometric difficulty: single-anchor replacement (L1), multi-anchor linking (L2), full-ring replacement (L3), and partial-ring modification (L4 Partial-ring).

Across L1–L3, RealFlux achieves Targeted Success values of 0.583–0.700 and P50 latency of 0.87–1.07 s, whereas DiffLinker exhibits P50 latency of 12.77–14.07 s. The L4 Partial-ring setting remains the principal failure boundary of the discrete editing strategy. Although RealFlux retains Chemical Validity of 0.950, Targeted Success falls to 0.100, and the companion predictor-representation diagnostic increases to $W = 3.1431$. Thus, high two-dimensional chemical validity alone does not guarantee successful or representation-proximal partial-ring counterfactuals (Table 7).

Table 7: Multi-level topological stress testing and execution-efficiency evaluation

Topological Stress Level	Method	Chemical Validity ↑	Context Retention ↑	Targeted Success Rate ↑	Synthetic Accessibility (SA) ↓	Latency P50 / P90 ↓ (s)
L1 (Single-anchor)	CM (CReM)	0.650	0.600	0.150	**2.893**	1.93 / 2.83
	CM (DiffLinker)	0.850	0.742	0.167	3.467	14.07 / 16.50
	RealFlux	**1.000**	**0.825**	**0.700**	3.343	**1.07 / 1.60**
L2 (Multi-anchor)	CM (CReM)	0.900	0.800	0.233	**2.930**	1.80 / 2.70
	CM (DiffLinker)	**0.950**	0.742	0.333	3.037	13.53 / 15.80
	RealFlux	0.900	**0.833**	**0.583**	3.683	**0.97 / 1.40**
L3 (Full-ring)	CM (CReM)	0.450	0.350	0.250	**3.117**	1.63 / 3.07
	CM (DiffLinker)	**0.883**	**0.658**	0.383	3.247	13.43 / 16.37
	RealFlux	0.850	0.588	**0.583**	3.417	**0.87 / 1.17**
L4 (Partial-ring)	CM (CReM)	0.000	0.000	0.000	N/A	2.03 / 4.10
	CM (DiffLinker)	0.850	**0.150**	0.000	N/A	12.77 / 15.43
	RealFlux	**0.950**	0.143	**0.100**	**3.810**	**0.90 / 1.07**

Note: Latency denotes the measured end-to-end wall-clock time for a single task, including candidate generation and forward property-validation overhead. All metrics are averaged across three random seeds.