MOLECULAR UNCERTAINTY GRADIENTS: EXACT DECOMPOSITION OF PREDICTIVE VARIANCE ONTO MOLECULAR SUBSTRUCTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Most explainability methods for molecular property prediction target predictions rather than uncertainty. Techniques for localizing predictive uncertainty to specific substructures remain comparatively underdeveloped, limiting our ability to identify which substructures drive uncertainty and which chemical patterns the model has not learned reliably. We present Molecular Uncertainty Gradients (MUG), a method that distributes a model's epistemic uncertainty, estimated as the variance across ensemble predictions, over the atoms of a molecule. By differentiating this variance with respect to intermediate atom representations, MUG computes signed atom-level contributions that sum exactly to the molecule's predictive variance. These attributions can be aggregated over any group of atoms using a single backward pass. MUG requires no retraining and works with the dropout-based uncertainty estimates already standard in practice. We show that MUG localizes uncertainty to molecular substructures absent from the training data. Across six molecular property prediction benchmarks, MUG achieves the highest occlusion-controlled fidelity, and its decomposition holds. Crucially, masking the features of its highest-ranked atoms reduces model uncertainty substantially more than masking other fragments. We further demonstrate that MUG's fragment-level uncertainty attributions can guide iterative library design toward compounds with favorable docking scores. Together, these results establish MUG as a complete and faithful approach to localizing predictive uncertainty, with practical applications in uncertainty-guided molecular design.

1 INTRODUCTION

Machine learning models for molecular property prediction are frequently applied to out-of-distribution molecules in tasks such as virtual screening, generative design, and multiparameter optimization. Standard uncertainty quantification (UQ) addresses this by providing a predictive variance alongside each prediction. While useful, a single molecule-level scalar is coarse: it signals that a prediction is uncertain, but cannot identify *which* substructure is responsible. Pinpointing the source of uncertainty is essential in practice—it tells medicinal chemists which R-groups to modify, helps data curators select informative fragments for training, and enables fragment-based active learning algorithms to select candidate substructures efficiently.

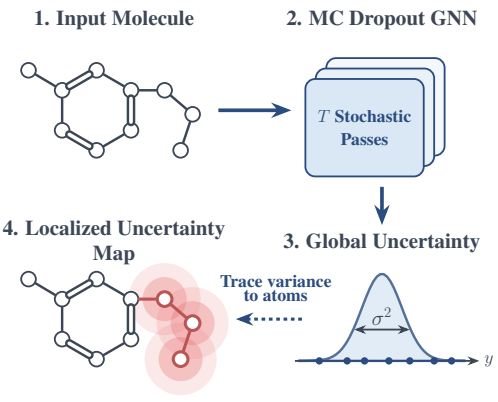


Figure 1: **Molecular Uncertainty Gradients (MUG).** Using backpropagated gradients with respect to predictive variance to pinpoint which chemical fragments cause model uncertainty.

Existing explainable AI (XAI) methods do not address this problem directly. Gradient-based attribution methods for graph neural networks (Pope et al., 2019; Simonyan et al., 2014; Selvaraju et al., 2019) explain *predictions* rather than uncertainties. Applying attribution to the predictive mean highlights features that drive the target output value, not

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.

those that cause ensemble *disagreement*; as we show empirically (§4.3), these two perspectives rank atoms differently. While Yang & Li (2023) introduced an atom-level uncertainty decomposition, their approach is restricted to a specific deep-ensemble architecture with dual mean-variance outputs, lacks mathematical completeness guarantees, and offers no quantitative ground-truth validation.

In this work, we present a principled attribution method based on a structural property of ensemble model estimators. The predictive variance of an ensemble is a differentiable function of its internal representations and is positively homogeneous of degree *two* with respect to the final-layer activations. Scaling these activations by $c > 0$ scales every member's prediction by c, scaling the ensemble variance by c^2. By Euler's homogeneous function theorem, calculating Grad $\times$ Input on the variance yields an exact mathematical identity rather than a heuristic approximation: the resulting attributions sum exactly to the predictive variance. Each substructure receives a signed share reflecting its contribution to model disagreement, providing an axiomatic completeness property analogous to Integrated Gradients and Shapley values (Sundararajan et al., 2017; Lundberg & Lee, 2017).

We name this method **Molecular Uncertainty Gradients** (MUG) and intuitively describe it in Figure 1. MUG requires no retraining, architectural modifications, or auxiliary output heads. It attributes predictive variance to molecular substructures using a single backward pass. Our key contributions are:

1. **Molecular Uncertainty Gradients (MUG)**, a method attributing predictive uncertainty to molecular substructures through signed gradient–activation products of predictive variance at the final node representations, yielding atom-level contributions in one backward pass.
2. **Theoretical guarantees**, establishing exact completeness after constant rescaling via the homogeneity of predictive variance and characterizing the readout and head assumptions under which this guarantee holds.
3. **Systematic empirical validation** across six molecular benchmarks and multiple GNN architectures. We validate our approach by removing (occluding) molecular fragments identified as high-uncertainty and measuring the resulting impact on model predictions and variance reduction. Finally, we demonstrate MUG's practical utility in drug discovery through simulated fragment-selection campaigns.

2 Related Work

Explaining graph neural networks. Graph neural networks (GNNs) (Bruna et al., 2014; Henaff et al., 2015) are widely used for molecular property prediction because molecules are naturally structured as graphs (Duvenaud et al., 2015; Gaudelet et al., 2021; Schapin et al., 2023). However, like most deep learning models, GNNs operate as black-box models. Gradient-based attribution originated in computer vision with saliency maps (Simonyan et al., 2014) and Grad-CAM (Selvaraju et al., 2019), which weight intermediate feature maps using backpropagated gradients, and was later extended to graphs by Pope et al. (2019). Another line of research trains auxiliary models to generate explanations (Ying et al., 2019; Luo et al., 2020); see Yuan et al. (2023) for a comprehensive taxonomy. Our work builds on *axiomatic* attribution methods, such as Integrated Gradients (Sundararajan et al., 2017), DeepLIFT (Shrikumar et al., 2017), and SHAP (Lundberg & Lee, 2017). These methods satisfy a completeness (or efficiency) axiom, guaranteeing that the attributions sum exactly to the explained quantity. However, all of these approaches explain model *predictions*. In contrast, we establish an equivalent completeness axiom for *predictive variance*, derived directly from the homogeneity of the variance functional without requiring path integrals or sampling schemes.

Uncertainty quantification. Bayesian neural networks (Goan & Fookes, 2020; Wang & Yeung, 2016; Kucukelbir et al., 2016) place probability distributions over model weights, but they incur high computational costs. More practical approximations dominate real-world applications, including deep ensembles (Lakshminarayanan et al., 2017), Monte-Carlo (MC) dropout (Gal & Ghahramani, 2016), evidential regression (Amini et al., 2020), and conformal prediction (Angelopoulos & Bates, 2022). These methods estimate aleatoric uncertainty (data noise) and epistemic uncertainty (model uncertainty due to limited training data coverage) (Kendall & Gal, 2017; Valdenegro-Toro & Saromo, 2022). In molecular property prediction, deep ensemble and MC dropout variance are the standard metrics for epistemic uncertainty and have been evaluated extensively (Hirschfeld et al., 2020; Scalia

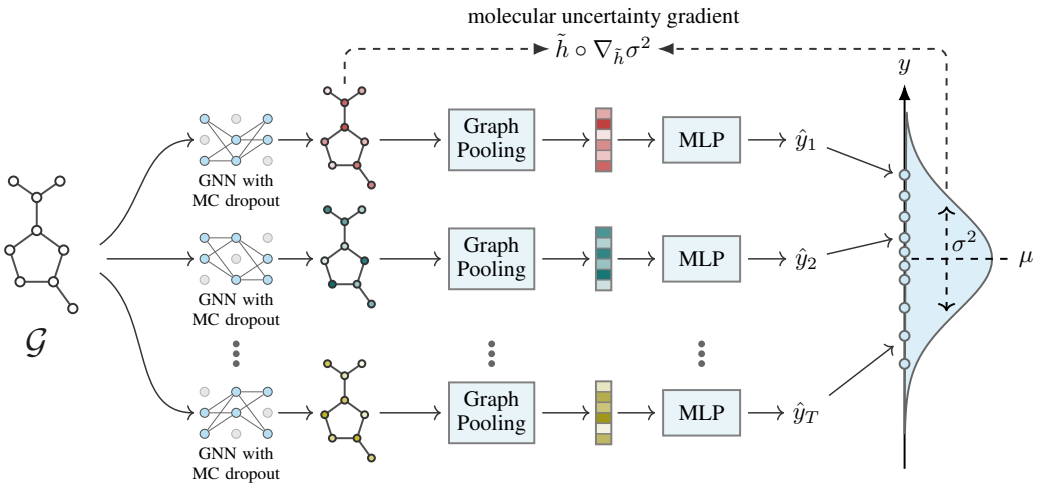


Figure 2: **Overview of the MUG framework.** An ensemble of T stochastic forward passes through a GNN with Monte-Carlo (MC) dropout maps an input molecular graph $\mathcal{G}$ to a set of predictions $\hat{y}_1, \ldots, \hat{y}_T$. The prediction variance σ^2 is backpropagated to the final graph layer to compute the molecular uncertainty gradient with respect to node representations $\tilde{h}$. Element-wise multiplication ($\tilde{h} \circ \nabla_{\tilde{h}} \sigma^2$) yields signed fragment attribution shares whose sum equals $\frac{2}{T}\sigma^2$.

et al., 2020). In this paper, we focus on epistemic uncertainty estimated via MC dropout and deep ensembles; however, MUG is agnostic to the underlying uncertainty estimator.

Explaining uncertainty. Few studies explain uncertainty at the atom level. Yang & Li (2023) decompose uncertainty with a deep ensemble whose members each predict a mean and a variance, so the method depends on this two-head architecture. AUINet (Qian et al., 2026) estimates atom-level uncertainty from the MC-dropout variance of atom representations and averages it into a molecular score. Neither method guarantees that the atom scores add up to the model's predictive variance. MUG works with any GNN that has a homogeneous readout, needs no retraining, and gives an exact decomposition.

3 Molecular Uncertainty Gradients

3.1 Preliminaries

Graph neural networks. A molecule is defined as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ with M nodes representing atoms, edges representing bonds, and node features $\mathbf{X} \in \mathbb{R}^{M \times d}$ encoding atomic properties. Message-passing networks update node representations over multiple layers. Let $m \in \{1, \ldots, M\}$ denote the atom index and $l \in \{1, \ldots, L\}$ denote the layer index. Starting from initial node features $h_m^{(0)} = x_m$, each layer updates the node state $h_m^{(l)}$ by aggregating messages from neighboring nodes. After L message-passing layers, a permutation-invariant readout function ρ aggregates the node states into a graph-level representation, which is then mapped to a scalar prediction $f(\mathcal{G})$ by a prediction head ψ: $f(\mathcal{G}) = \psi\Big(\rho\Big(\{h_m^{(L)}\}_{m=1}^{M}\Big)\Big)$.

Monte-Carlo dropout. Epistemic uncertainty can be estimated using ensembles. Monte-Carlo dropout (Gal & Ghahramani, 2016) constructs an implicit ensemble from a single network by maintaining active dropout at inference, resulting in sampling T stochastic forward passes $\{f_t\}_{t=1}^{T}$. The predictive mean $\mu(\mathcal{G})$ and predictive variance $\sigma^2(\mathcal{G})$ across these passes are defined as:

$$\mu(\mathcal{G}) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathcal{G}), \qquad \sigma^2(\mathcal{G}) = \frac{1}{T-1}\sum_{t=1}^{T}\big(f_t(\mathcal{G}) - \mu(\mathcal{G})\big)^2. \tag{1}$$

The ensemble need not be implicit. The set $\{f_t\}_{t=1}^{T}$ can equally represent T independently trained networks, provided they share a readout bias. MC dropout satisfies this condition automatically because all stochastic passes share a single output head (Proposition 1).

Gradient-based explanations. Grad-CAM (Selvaraju et al., 2019; Pope et al., 2019) computes node importance by backpropagating the gradient of a scalar model prediction $\hat{y}$ to an intermediate layer and multiplying it by the corresponding layer activations. Let $\tilde{h}_{mk}$ denote the k-th feature channel of node m's representation in the final layer (and $\tilde{h}_{mk}^{(t)}$ for pass t). Grad-CAM computes node attribution E_m as:

$$\alpha_k = \frac{1}{M} \sum_{m=1}^{M} \frac{\partial \hat{y}}{\partial \tilde{h}_{mk}}, \qquad E_m = \sum_{k=1}^{K} \alpha_k \, \tilde{h}_{mk}, \tag{2}$$

where K is the number of feature channels. Standard implementations apply a Rectified Linear Unit (ReLU) to E_m to keep only positive evidence. While effective for prediction attribution, we demonstrate that this rectification destroys the completeness property when explaining uncertainty.

3.2 Attributing the Variance

We propose how gradients can be used to explain model uncertainty. MUG differentiates the *predictive variance* $\sigma^2(\mathcal{G})$ instead of a single prediction (Figure 2). Let

$$\varsigma_{mk}^{(t)} \;=\; \frac{\partial\, \sigma^2(\mathcal{G})}{\partial\, \tilde{h}_{mk}^{(t)}}, \qquad E_m^{\sigma^2} \;=\; \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} \varsigma_{mk}^{(t)} \, \tilde{h}_{mk}^{(t)}. \tag{3}$$

Because each stochastic forward pass t yields distinct activations $\tilde{h}^{(t)}$, gradients are computed with respect to T separate feature tensors and averaged across passes. Two key mathematical properties follow directly from this formulation (proven in Appendices A and B).

Lemma 1 (Deviation proportionality)**.** *With $\sigma^2(\mathcal{G})$ defined as in Eq. 1, $\varsigma_{mk}^{(t)} = \frac{2}{T-1}\big(f_t(\mathcal{G}) - \mu(\mathcal{G})\big)\,\frac{\partial f_t(\mathcal{G})}{\partial \tilde{h}_{mk}^{(t)}}$. Consequently, $E_m^{\sigma^2} = \sum_{t=1}^{T} c_t$ with $c_t \propto f_t(\mathcal{G}) - \mu(\mathcal{G})$: each member's contribution is weighted by its deviation from the mean prediction, so members agreeing with the consensus contribute zero.*

Corollary 1 (Zero uncertainty, zero attribution)**.** *If $\sigma^2(\mathcal{G}) = 0$ then $E_m^{\sigma^2} = 0$ for every m.*

Proof. By Lemma 1 every $\varsigma_{mk}^{(t)}$ carries the factor $f_t(\mathcal{G}) - \mu(\mathcal{G})$, and $\sigma^2(\mathcal{G}) = 0$ forces every such factor to vanish. No assumption on ρ or ψ is needed, so the corollary holds even where Theorem 1 below does not. □

Theorem 1 (Completeness)**.** *Assume the composition of readout ρ and head ψ is positively homogeneous of degree one in the final-layer activations (e.g., mean- or sum-pooling followed by a linear head or a bias-free ReLU MLP), and that the prediction head is shared across all T ensemble members. Then the following equation holds exactly:*

$$\sum_{m=1}^{M} E_m^{\sigma^2} \;=\; \frac{2}{T}\, \sigma^2(\mathcal{G}). \tag{4}$$

Theorem 1 establishes MUG as an exact mathematical decomposition. We define the *normalized share* of fragment m as:

$$\boxed{\hat{E}_m \;=\; \frac{E_m^{\sigma^2}}{\sum_{m'} E_{m'}^{\sigma^2}} \;=\; \frac{T\, E_m^{\sigma^2}}{2\, \sigma^2(\mathcal{G})}}, \tag{5}$$

which represents the exact fraction of total predictive variance assigned to fragment m. This normalized share is invariant to the constant $2/T$, pooling normalization, and the variance estimator convention (biased vs. unbiased; see Appendix B). Crucially, these shares are *signed*: a negative share $E_m^{\sigma^2} < 0$ indicates that fragment m reduces ensemble disagreement and stabilizes the model's prediction.

Proposition 1 (Bias invariance and ensemble-bias correction)**.** *Let member t have a linear head $\psi_t(z) = w_t^\top z + b_t$, and let $\mathrm{Cov}_t(u_t, v_t) = \frac{1}{T-1}\sum_{t=1}^{T}(u_t - \bar{u})(v_t - \bar{v})$ denote the sample covariance over members, the convention of Eq. 1. Then:*

$$\sum_{m=1}^{M} E_m^{\sigma^2} \;=\; \frac{2}{T}\,\sigma^2(\mathcal{G}) \;-\; \frac{2}{T}\,\mathrm{Cov}_t(f_t(\mathcal{G}),\, b_t)\,. \tag{6}$$

Thus, identity (4) *holds exactly whenever a member's head bias is uncorrelated with its prediction, specifically when the head bias is shared ($b_t \equiv b$, as in MC dropout) or zero.*

Proposition 1 (proof in Appendix C) states that a bias in the head does no harm as long as every member has the same one. The variance only sees differences between members, and a shared bias drops out of every difference. Monte-Carlo dropout is exactly this case, since all passes share one head. Prediction attribution is less forgiving: Grad $\times$ Input on a single prediction sums to $f - b$, so any bias breaks it. In a deep ensemble, each member trains its own head, so its bias moves with its prediction and the attributions fall short by the covariance term in Eq. 6. That term is cheap to compute and can be subtracted, or the biases can be tied or removed when the ensemble is trained.

3.3 Scope of the Guarantee

The completeness guarantee holds for a specific class of network architectures outlined below.

A1 (Homogeneous readout). ρ is mean-, sum-, or max-pooling. Softmax attention pooling violates degree-1 homogeneity.
A2 (Homogeneous head). ψ is linear (shared biases permitted via Prop. 1) or a bias-free ReLU MLP. Internal biases, BatchNorm running statistics, and non-homogeneous activations (sigmoid, tanh, GELU) violate degree-1 homogeneity.
A3 (Signed attribution). $E_m^{\sigma^2}$ is unrectified, as defined in Eq. 3 (§F.8).
A4 (Logit space). For classification tasks, σ^2 is computed on pre-sigmoid logits. Under delta-method linearization, normalized shares $\hat{E}_m$ are preserved, though the scaling constant changes.
A5 (Shared readout bias). Ensemble members share a head or use bias-free heads (satisfied by MC dropout). Deep ensembles with independent heads incur the closed-form correction in Eq. 6.

4 Experiments

Evaluating uncertainty attributions is challenging because no ground-truth atom-level decomposition of predictive uncertainty exists. We therefore evaluate MUG through complementary tests of **sensitivity to chemical novelty** (§4.1,§4.4), **faithfulness** (§4.2), **specificity** (§4.3), and **completeness** (§4.5), before demonstrating its **utility for molecular design** (§5).

Experimental setup. We evaluate MUG on six Therapeutics Data Commons benchmarks (Huang et al., 2021): three regression tasks (Caco2, Lipophilicity, and AqSolDB) and three binary classification tasks (BBBP, Tox21 NR-ER, and CYP3A4). Unless stated otherwise, we use $70\%/10\%/20\%$ train/validation/test Bemis–Murcko scaffold splits (Bemis & Murcko, 1996), a GCN backbone, five model seeds, and $T = 800$ MC dropout samples. We additionally test completeness with GIN (Xu et al., 2019), D-MPNN (Yang et al., 2019), and R-MAT (Maziarka et al., 2021). Dataset references, experimental details, and sensitivity to T are shown in Appendix D.

Evaluation metrics. In our experiments, we use four core metrics to assess the completeness and faithfulness of uncertainty attributions. **(i) Relative completeness residual** measures the relative discrepancy between summed atom-level attributions and the scaled molecule-level predictive variance; lower values indicate better completeness. **(ii) Top-ranked occlusion fidelity** ($\mathrm{Fid}^{\downarrow}$) measures the relative reduction in predictive uncertainty after masking the highest-ranked atoms. **(iii) Controlled fidelity** ($\Delta\mathrm{Fid} = \mathrm{Fid}^{\downarrow} - \mathrm{Fid}^{\uparrow}$) compares this reduction with that obtained by masking the same number of lowest-ranked atoms. **(iv) Selectivity rate** measures how often top-ranked masking yields lower uncertainty than bottom-ranked masking. Formal definitions and masking details are provided in Appendix E; experiment-specific metrics are introduced where used.

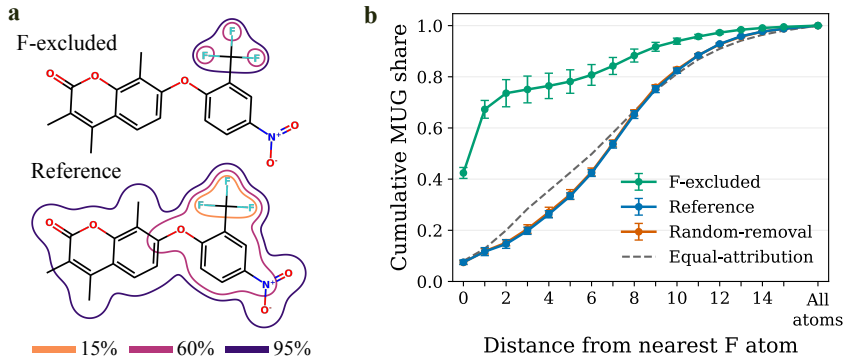


Figure 3: **Withholding fluorinated training molecules localizes uncertainty attribution near fluorine across CYP3A4 test molecules.** **(a)** Attribution for a single test molecule. Contours enclose 15%, 60%, and 95% of total MUG attribution. The F-excluded model has a substantially higher predictive uncertainty than the reference model (1.40 vs. 0.32) and concentrates its uncertainty attribution around the trifluoromethyl group. **(b)** Cumulative MUG attribution within r bonds of fluorine across fluorine-containing test molecules. The F-excluded model concentrates attribution near fluorine, whereas the reference and random-removal models have nearly identical profiles close to the equal-per-atom baseline. Mean $\pm$ SD over five independently trained seeds.

4.1 What the model has not seen: localized uncertainty around unseen atoms

We investigate whether MUG localizes uncertainty around an atom type absent from the training data. On CYP3A4, we compare an *F-excluded model* with a *reference model* trained on the complete training set. We also include a *random-removal control*, for which the same number of training molecules is removed at random, to account for the reduction in training-set size. All three models are evaluated on the same fluorine-containing test molecules.

For a single test molecule, Figure 3a shows the region containing 95% of total MUG attribution concentrated around the trifluoromethyl group under the F-excluded model, but spread across the molecule under the reference model. This localization is accompanied also by an increase in predictive uncertainty from 0.32 to 1.40.

Figure 3b summarizes cumulative MUG attribution by topological distance from the nearest fluorine atom across fluorine-containing CYP3A4 test molecules. The F-excluded model assigns approximately 42% of attribution directly to fluorine and two thirds within one bond, compared with only 8% directly to fluorine for the reference and random-removal models, whose profiles nearly overlap the equal-per-atom baseline. These results show that, on average, MUG attribution is more concentrated near the excluded atom type.

Appendix Fig. 7 further shows that fluorine exclusion increases predictive uncertainty for fluorinated test molecules, with the strongest shift for molecules containing three or more fluorine atoms. We also observe similar localization of uncertainty for larger functional groups, such as sulfone, when molecules containing the group are excluded from training (Appendix F.3).

4.2 Occlusion Fidelity of Uncertainty Attributions

If an attribution is faithful, masking the atoms it ranks highest should lower the predictive uncertainty more than masking the atoms it ranks lowest. For each test molecule, we rank atoms by their signed score, zero the input features of the top and the bottom $k = \lceil 0.1M \rceil$ atoms in two separate runs, and recompute σ without retraining. Masking any atom moves the molecule away from the training data and can change σ on its own. We therefore focus on $\Delta\mathrm{Fid}$, which compares two masks of the same size, and report $\mathrm{Fid}^{\downarrow}$ and the selectivity rate next to it.

We compare MUG with a random and an atom-degree ranking, with Saliency and GNNExplainer applied to the same MC-dropout variance, and with two atom-level uncertainty methods, Yang & Li (Yang & Li, 2023) and AUINet (Qian et al., 2026). The last two use their own uncertainty estimators, so for them we test estimator and attribution together. Yang & Li (2023) supports regression only.

Table 1: **Occlusion fidelity across regression and classification tasks.** Top- and bottom-ranked atoms are masked separately using $k = \lceil 0.1M \rceil$. Results are three-task macro-averages within regression and classification (mean ± SD over five seeds). Best means are bold and second-best means are underlined.

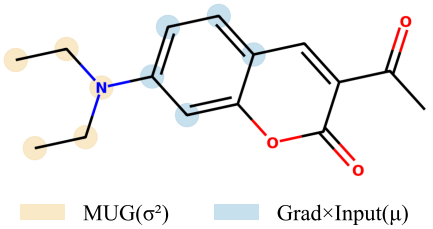
	Regression			Classification		
Atom attribution	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Selectivity rate (%)↑	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Selectivity rate (%)↑
Random	-0.001 ± 0.005	0.000 ± 0.021	50.21 ± 0.89	0.004 ± 0.004	-0.033 ± 0.016	50.70 ± 0.76
Degree	0.183 ± 0.123	0.016 ± 0.094	74.55 ± 9.44	0.058 ± 0.056	-0.078 ± 0.047	57.31 ± 4.83
Saliency	$\underline{0.228 \pm 0.028}$	$\underline{0.137 \pm 0.025}$	$\underline{89.73 \pm 2.70}$	$\underline{0.168 \pm 0.019}$	$\mathbf{0.103 \pm 0.014}$	$\mathbf{83.29 \pm 1.57}$
GNNExplainer	0.016 ± 0.007	0.008 ± 0.022	55.10 ± 1.83	0.016 ± 0.005	-0.025 ± 0.015	52.07 ± 1.17
Yang & Li	0.085 ± 0.014	0.005 ± 0.039	80.21 ± 5.10	–	–	–
AUINet	0.000 ± 0.004	0.031 ± 0.005	53.38 ± 3.24	-0.002 ± 0.004	0.036 ± 0.004	53.05 ± 5.26
MUG	$\mathbf{0.229 \pm 0.037}$	$\mathbf{0.145 \pm 0.034}$	$\mathbf{90.23 \pm 4.35}$	$\mathbf{0.172 \pm 0.019}$	$\underline{0.101 \pm 0.021}$	$\underline{82.25 \pm 1.69}$

Table 2: **Prediction and uncertainty attributions rank atoms differently.** Grad×Input at each layer is compared with final-layer MUG(σ^2). Spearman's ρ uses absolute atom-level attributions; top-atom disagreement denotes disjoint sets of top-ranked atoms, including ties. Results are six-task macro-averages (mean ± SD over five seeds).

Prediction attribution	Mean $\rho \downarrow$	Top disagreement ↑
Final GNN layer	0.403 ± 0.024	$(70.8 \pm 0.6)\%$
Input layer	0.376 ± 0.012	$(80.2 \pm 1.2)\%$

MUG has the highest ΔFid in Table 1 (0.229 for regression, 0.172 for classification). Masking its top atoms lowers the uncertainty by 14.5% and 10.1%, and the top mask beats the bottom mask in 90% and 82% of molecules. Existing uncertainty attributions do much worse. AUINet and GNNExplainer stay close to the random rate of 50%. Yang & Li reaches 80% selectivity, but its $\text{Fid}^{\downarrow}$ is near zero: masking its top atoms barely changes the uncertainty, and the gap comes from its bottom atoms raising it. The degree ranking behaves the same way and fails on CYP3A4 ($\Delta\text{Fid} = -0.241$), while MUG is positive on all six datasets (Appendix F.4).

Saliency scores about as well as MUG. This is expected, since both differentiate the same variance. Together with §4.3, it shows that the target of the gradient matters more than the attribution rule. What MUG adds is an exact, signed decomposition: its completeness residual is about 10^{-5}, compared with about 10^{3} for Saliency (Appendix F.1).



Attribution	ΔFid_{μ}	$\Delta\text{Fid}_{\sigma}$
MUG(σ^2)	-0.333 ± 0.227	$\mathbf{+0.355 \pm 0.160}$
Grad×Input(μ)	$\mathbf{+0.372 \pm 0.312}$	-0.400 ± 0.077

Figure 4: **Prediction and uncertainty highlight distinct substructures.** Final-layer MUG highlights the N,N-diethylamino substituent (orange), while Grad×Input highlights the aromatic core (blue). Atoms are ranked by absolute mean attribution across seeds. Fidelity uses $k = 5$ masked atoms (mean ± SD over five seeds).

4.3 Uncertainty Attribution Is Not Prediction Attribution

A natural concern is that uncertainty attribution may simply reproduce prediction attribution. We compare MUG, which attributes the predictive variance σ^2, with Grad×Input, which attributes the predictive mean μ. We compute both methods at the final GNN layer, using the same molecular representation to isolate the effect of the attribution target. We also evaluate input-layer Grad×Input as a conventional prediction-attribution baseline.

Figure 4 shows the distinction for a Lipophilicity test molecule. Grad×Input(μ) highlights the conjugated aromatic core (blue), whereas MUG(σ^2) identifies the N,N-diethylamino substituent (orange). We found that this substituent's direct attachment to an aromatic ring is absent from the training set. The two methods' opposing fidelity scores support this distinction: Grad×Input is faithful to the predictive mean, whereas MUG is faithful to predictive uncertainty.

Table 3: **Fidelity depends on the attribution target.** We mask the top- and bottom-ranked 10% of atoms ($k = \lceil 0.1M \rceil$) and evaluate fidelity to the predictive mean ($\Delta\mathrm{Fid}_\mu$) and uncertainty ($\Delta\mathrm{Fid}_\sigma$). Results are six-task macro-averages (mean $\pm$ SD over five seeds).

Attribution	Prediction $\Delta\mathrm{Fid}_\mu$	Uncertainty $\Delta\mathrm{Fid}_\sigma$
Grad$\times$Input(μ), final	$+0.259 \pm 0.005$	-0.100 ± 0.015
Grad$\times$Input(μ), input	$\mathbf{+0.417 \pm 0.023}$	-0.075 ± 0.029
MUG(σ^2)	$+0.209 \pm 0.025$	$\mathbf{+0.199 \pm 0.026}$

We next compare the two attribution targets across six tasks. For each molecule, we compute Spearman's rank correlation between absolute atom-level attributions. Table 2 shows partial rank agreement and frequent top-atom disagreement. Table 3 further shows that each method achieves higher fidelity to its own attribution target. On average across tasks, Grad$\times$Input achieves positive prediction fidelity at both layers but negative uncertainty fidelity. MUG achieves positive uncertainty fidelity, while its prediction fidelity is positive but lower than that of both prediction baselines. Together, these results show that prediction attribution does not substitute for uncertainty attribution in these experiments. Comparison details and dataset-level results are provided in Appendix F.5.

4.4 MUG Highlights Training-Unseen Substructures

Motivated by the hypothesis that rarer structural features receive larger uncertainty attributions (§4.3), and supported by the results of the controlled atom-type withholding experiment, we next ask whether the same pattern arises under standard dataset splits. We decompose molecules into BRICS fragments and classify each test fragment as seen or unseen by exact canonical SMILES matching against training fragments. Across six datasets, unseen fragments receive 3.7–6.6 times higher mean attributions of uncertainty. Although unseen fragments are larger, their mean per-atom attributions remain 1.3–4.6 times higher after dividing each fragment's attribution by its atom count. These results support an association between exact-fragment novelty and higher MUG attribution beyond simple size accumulation. Appendix F.6 reports the full results and shows that elevated MUG uncertainty can also occur in compounds similar to those in the training set.

Table 4: **Unseen fragments receive higher attributions.** Unseen-to-seen ratios of means across five seeds.

Dataset	MUG	MUG/atom
BBBP	6.6	4.6
CYP3A4	4.0	1.5
Caco2	4.4	1.3
Lipophilicity	3.7	1.5
AqSolDB	5.7	1.6
Tox21	5.9	2.4

4.5 Scope and Limits of the Completeness Guarantee

We test the scope of Theorem 1 across four GNN architectures. GCN, GIN, and D-MPNN configurations satisfying its homogeneity assumptions yield completeness residuals of 10^{-6}–10^{-5}, compared with approximately 0.17 for R-MAT with nonhomogeneous attention pooling (Table 5).

Table 5: **Completeness across architectures.** Relative completeness residual and occlusion fidelity, averaged across six datasets.

Backbone	Pooling	Residual $\downarrow$	ΔFid $\uparrow$
GCN	Mean	1.10×10^{-5}	0.182
GIN	Mean	1.17×10^{-5}	0.194
D-MPNN	Sum	3.28×10^{-6}	0.157
R-MAT	Attention	1.69×10^{-1}	-0.024

Controlled GCN ablations further isolate these limits: attention pooling, biased ReLU heads, and input-level attribution produce larger residuals, whereas changing the MC-dropout sample count or variance normalization preserves completeness (Appendix F.7). Preserving signed contributions is also essential: applying ReLU breaks completeness whenever negative attributions are present (Appendix F.8).

5 Case Study: Uncertainty-Guided Fragment Library Design

Combinatorial libraries are built from reusable building blocks. When a model is unsure about a compound, MUG tells us which building block causes the doubt, and we can test more products of that block. We check whether this helps in a simulated active-learning campaign.

Table 6: **Uncertainty-guided library design for 5L2S.** 50 initial compounds, followed by 50 rounds of 25 acquisitions. Hits score ≤ -9.193 kcal/mol (top 10% of the docked library). Results are mean $\pm$ SD over five seeds. Best means are bold and second-best means are underlined.

Acquisition	Avg. score (kcal/mol) ↓	Best score (kcal/mol) ↓	Hit rate ↑
Random molecules	-8.326 ± 0.021	-10.922 ± 0.363	0.109 ± 0.007
Greedy	-9.123 ± 0.010	-11.404 ± 0.104	0.454 ± 0.010
UCB	-9.106 ± 0.018	-11.366 ± 0.104	0.442 ± 0.009
Random fragments	-8.324 ± 0.035	-10.958 ± 0.256	0.100 ± 0.014
MUG (random fragment)	-9.134 ± 0.018	$\mathbf{-11.480 \pm 0.000}$	0.472 ± 0.010
MUG	$\mathbf{-9.283 \pm 0.027}$	$\mathbf{-11.480 \pm 0.000}$	$\mathbf{0.562 \pm 0.019}$

Setup. The library contains 13,176 acid–amine products from 91 acids and 120 amines of Enamine's Essential Fragment Library (Enamine Ltd., 2024). We dock every product against human CDK6 (PDB `5L2S`; Chen et al., 2016) with AutoDock Vina (Eberhardt et al., 2021) and call a compound a hit if it is in the top 10% of the library (≤ -9.193 kcal/mol). Each campaign starts from 50 labelled compounds and runs 50 rounds of 25 acquisitions, over five seeds. In every round we retrain a three-layer GCN, score 500 random unlabelled candidates, and select a batch.

Acquisition rules. Greedy selects the lowest predicted scores μ. UCB selects the lowest $\mu - \sigma$, with σ from MC dropout. MUG ranks candidates by two criteria with equal weight: a low μ and a large MUG share on one of the two building blocks. For each top candidate, it takes the block with the larger share and acquires up to five other products of that block from the whole pool, choosing the ones with the lowest μ. In the MUG (random fragment) ablation, the block is picked by a coin flip. Random fragments also picks the candidates at random, and Random molecules samples compounds uniformly. Details are in Appendix F.9.

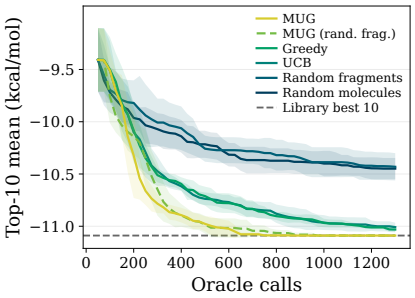


Figure 5: **Mean docking score of the 10 best compounds acquired so far for 5L2S (lower is better).** The gray dashed line marks the library's top-10 mean.

Results. MUG finds the most hits (Table 6). Its hit rate is 56.2%, against 45.4% for greedy and 10.9% for random sampling. That is about 700 hits instead of 570 from the same 1,250 docking calls, and MUG matches greedy's final count after about 1,050 calls. It also finds good compounds sooner. The mean of its ten best compounds reaches that of the whole library after about 650 calls, while greedy and UCB still fall short after 1,250 (Fig. 5). On a second target, `8PJK`, MUG reaches 52.3% hits against 42.8% for greedy (Appendix F.9).

6 Conclusion

MUG connects molecule-level uncertainty estimates to the substructures that contribute to them. Using a single backward pass without retraining, it assigns signed atom-level contributions that can be aggregated into fragments and, under the stated architectural assumptions, rescaled to sum exactly to predictive variance. Across six molecular property benchmarks, MUG identifies substructures absent from training and achieves the highest occlusion-controlled fidelity among evaluated methods, revealing chemical patterns the model has not learned reliably. In a simulated library-design study, attribution-guided fragment selection increases the docking hit rate from 47.2% to 56.2% compared with random fragment selection under the same acquisition framework, demonstrating its potential for uncertainty-guided molecular design.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our theoretical claims and empirical findings, we take the following measures:

- **Theoretical Results:** Complete, step-by-step mathematical proofs for all theoretical guarantees, including Theorem 1 regarding the exact summation of molecular uncertainty shares to $\frac{2}{T}\sigma^2(\mathcal{G})$, are provided in Appendix B.
- **Source Code & Environment:** Anonymized source code implementing MUG, along with environment specifications, training scripts, and demo Jupyter notebooks, is submitted in the supplementary material. Upon publication, the full repository will be made publicly available on GitHub.
- **Datasets & Preprocessing:** All experiments are conducted on standard, publicly available molecular benchmarks from Therapeutics Data Commons. Data splitting protocols, canonicalization steps (via RDKit), and filtering criteria are detailed in Appendix D.4 and D.6.
- **Hyperparameters & Training Details:** A comprehensive list of hyperparameters, including GNN architecture choices, learning rates, MC dropout rates, the number of stochastic passes T, random seeds, and compute hardware, is documented in Appendix D.

ETHICS STATEMENT

This work uses public, non-human molecular property datasets and involves no human subjects. Its intended effect is to make failures of molecular property models legible, which supports rather than undermines safe deployment. We note the dual-use character common to all molecular property prediction: an acquisition function that designs a focused library toward a therapeutic target could in principle be pointed at a harmful one. The method confers no capability that molecule-level uncertainty and standard docking do not already confer, and we release no models trained on toxicity data for offensive use. The docking study in §5 is computational; docking scores are not affinity measurements and none of the designed compounds were synthesized.

LLM USAGE STATEMENT

In accordance with ICLR guidelines, we explicitly disclose the scope and limits of Large Language Model (LLM) assistance during the preparation of this work:

- **Writing and Copyediting (Used):** LLMs (specifically ChatGPT and Claude) were used as copyediting tools to refine phrasing, correct grammar, and improve the overall clarity of author-written text. All generated suggestions were manually reviewed and approved by the authors.
- **Literature Discovery (Used):** LLMs were consulted during initial literature retrieval to identify relevant prior work on uncertainty estimation and graph neural networks. Every cited reference was independently read, verified, and contextualized by the authors.
- **Mathematical Proof Assistance (Used):** LLMs were consulted to sanity-check intermediate algebraic steps and verify proof readability for theoretical claims (including Theorem 1).
- **Research Execution and Coding (Used):** LLMs were utilized as coding assistants for debugging code snippets, refactoring functions, and troubleshooting environment configurations during the pipeline implementation. The overall architectural logic, experimental execution, and data analysis were directed and validated entirely by the authors.
- **Research Ideation and Conceptualization (Not Used):** The core research questions, theoretical formulations, and methodological design of MUG were conceived entirely by the authors without LLM involvement.

The authors assume full responsibility for the accuracy, novelty, and integrity of all text, mathematical proofs, code, and empirical results presented in this manuscript.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2022.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs, 2014.

Ping Chen, Nathan V Lee, Wenyue Hu, Meirong Xu, Rose Ann Ferre, Hieu Lam, Simon Bergqvist, James Solowiej, Wade Diehl, You-Ai He, et al. Spectrum and degree of cdk drug interactions predicts clinical performance. *Molecular cancer therapeutics*, 15(10):2273–2281, 2016.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints, 2015.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

Enamine Ltd. Essential fragment library. `https://enamine.net/compound-libraries/fragment-libraries/essential-fragment-library`, 2024. Accessed 2026-07-10.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016.

Thomas Gaudelet, Ben Day, Arian R. Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy B. R. Hayter, Richard Vickers, Charles Roberts, Jian Tang, David Roblin, Tom L. Blundell, Michael M. Bronstein, and Jake P. Taylor-King. Utilising graph machine learning within drug discovery and development, 2021.

Ethan Goan and Clinton Fookes. *Bayesian Neural Networks: An Introduction and Survey*, pp. 45–87. Springer International Publishing, 2020. ISBN 9783030425531. doi: 10.1007/978-3-030-42553-1_3. URL `http://dx.doi.org/10.1007/978-3-030-42553-1_3`.

Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data, 2015.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020.

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 *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Frontiers in Environmental Science*, 3:85, 2016.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference, 2016. URL `https://arxiv.org/abs/1603.00788`.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2017. URL https://arxiv.org/abs/1612.01474.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling*, 52(6):1686–1697, 2012.

Łukasz Maziarka, Dawid Majchrowski, Tomasz Danel, Piotr Gaiński, Jacek Tabor, Igor Podolak, Paweł Morkisz, and Stanisław Jastrzębski. Relative molecule self-attention transformer, 2021.

Phillip E. Pope, Soheil Kolouri, Mohammad Rostami, Charles E. Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10764–10773, 2019. doi: 10.1109/CVPR.2019.01103.

Jiayu Qian, Yukun Luo, Qingping Zhou, and Xiaoqing Guan. Atomic uncertainty pinpoints critical failure structures for trustworthy molecular property prediction. *Journal of Chemical Information and Modeling*, 2026.

Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, 2020.

Nikolai Schapin, Maciej Majewski, Alejandro Varela, Carlos Arroniz, and Gianni De Fabritiis. Machine learning small molecule properties in drug discovery, 2023.

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. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

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(1):143, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3319–3328. PMLR, 2017.

Matias Valdenegro-Toro and Daniel Saromo. A deeper look into aleatoric and epistemic uncertainty disentanglement, 2022.

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.

Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, 2016. doi: 10.1109/TKDE.2016.2606428.

Ning-Ning Wang, Jie Dong, Yin-Hua Deng, Min-Feng Zhu, Ming Wen, Zhi-Jiang Yao, Ai-Ping Lu, Jian-Bing Wang, and Dong-Sheng Cao. ADME properties evaluation in drug discovery: Prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting. *Journal of Chemical Information and Modeling*, 56(4):763–773, 2016.

Mark Wenlock and Nicholas Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. *ChEMBL*, 2015.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019. URL https://arxiv.org/abs/1810.00826.

Chu-I Yang and Yi-Pei Li. Explainable uncertainty quantifications for deep learning-based molecular property prediction. *Journal of Cheminformatics*, 15(1):13, 2023.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8): 3370–3388, 2019.

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, 2019.

Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5782–5799, 2023.

A Proof of Lemma 1

Proof. The activations $\tilde{h}^{(j)}$ of pass j influence σ^2 only through f_j and, via f_j, through μ. Differentiating Eq. 1,

$$\frac{\partial \sigma^2(\mathcal{G})}{\partial \tilde{h}^{(j)}_{mk}} = \frac{2}{T-1}\sum_{t=1}^{T}\big(f_t(\mathcal{G})-\mu(\mathcal{G})\big)\left(\frac{\partial f_t(\mathcal{G})}{\partial \tilde{h}^{(j)}_{mk}} - \frac{\partial \mu(\mathcal{G})}{\partial \tilde{h}^{(j)}_{mk}}\right). \tag{7}$$

Since $\partial f_t/\partial \tilde{h}^{(j)}_{mk} = 0$ for $t \neq j$ and $\partial \mu/\partial \tilde{h}^{(j)}_{mk} = \frac{1}{T}\,\partial f_j/\partial \tilde{h}^{(j)}_{mk}$,

$$\frac{\partial \sigma^2(\mathcal{G})}{\partial \tilde{h}^{(j)}_{mk}} = \frac{2}{T-1}\left[\big(f_j-\mu\big)\frac{\partial f_j}{\partial \tilde{h}^{(j)}_{mk}} - \frac{1}{T}\frac{\partial f_j}{\partial \tilde{h}^{(j)}_{mk}}\sum_{t=1}^{T}\big(f_t-\mu\big)\right] = \frac{2}{T-1}\big(f_j-\mu\big)\frac{\partial f_j}{\partial \tilde{h}^{(j)}_{mk}}, \tag{8}$$

because $\sum_t (f_t - \mu) = 0$. Substituting into Eq. 3,

$$E_m^{\sigma^2} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\varsigma^{(t)}_{mk}\,\tilde{h}^{(t)}_{mk} = \sum_{t=1}^{T}\underbrace{\frac{2}{T(T-1)}}_{\text{constant}}\underbrace{\big(f_t(\mathcal{G})-\mu(\mathcal{G})\big)}_{\text{deviation}}\sum_{k=1}^{K}\frac{\partial f_t(\mathcal{G})}{\partial \tilde{h}^{(t)}_{mk}}\,\tilde{h}^{(t)}_{mk}, \tag{9}$$

which exhibits $E_m^{\sigma^2}$ as $\sum_t c_t$ with $c_t \propto f_t - \mu$. □

B Proof of Theorem 1

We give both the direct derivation, which yields the constant, and the Euler argument, which explains it.

Direct derivation. Assume mean-pooling and a linear head shared across the ensemble members (assumption A5; Appendix C treats per-member heads), $f_t(\mathcal{G}) = w^\top \big(\frac{1}{M}\sum_m \tilde{h}_m^{(t)}\big) + b$, so that $\partial f_t / \partial \tilde{h}_{mk}^{(t)} = w_k / M$. Write $p_m^{(t)} := \frac{1}{M}\sum_k w_k \tilde{h}_{mk}^{(t)}$ for the contribution of fragment m to pass t, whence $\sum_m p_m^{(t)} = f_t - b$. By Lemma 1,

$$E_m^{\sigma^2} = \frac{2}{T(T-1)} \sum_{t=1}^{T} \big(f_t - \mu\big)\, p_m^{(t)}. \tag{10}$$

Summing over fragments,

$$\sum_{m=1}^{M} E_m^{\sigma^2} = \frac{2}{T(T-1)} \sum_{t=1}^{T} \big(f_t - \mu\big)\big(f_t - b\big) \tag{11}$$

$$= \frac{2}{T(T-1)} \left[\sum_{t=1}^{T} \big(f_t - \mu\big)^2 + (\mu - b) \underbrace{\sum_{t=1}^{T} \big(f_t - \mu\big)}_{=\,0} \right] \tag{12}$$

$$= \frac{2}{T(T-1)}\,(T-1)\,\sigma^2(\mathcal{G}) \;=\; \frac{2}{T}\,\sigma^2(\mathcal{G}). \qquad \blacksquare \tag{13}$$

Euler argument. Let $\Phi = \{\tilde{h}_{mk}^{(t)}\}$ denote all final-layer activations across passes. If $\psi \circ \rho$ is positively homogeneous of degree one, then $f_t(c\,\Phi) = c\, f_t(\Phi)$ for $c > 0$, hence $\mu(c\Phi) = c\,\mu(\Phi)$ and $\sigma^2(c\Phi) = c^2\,\sigma^2(\Phi)$. Euler's homogeneous function theorem applied to σ^2, which is homogeneous of degree two, gives

$$\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{k=1}^{K} \tilde{h}_{mk}^{(t)}\, \frac{\partial \sigma^2}{\partial \tilde{h}_{mk}^{(t)}} \;=\; 2\,\sigma^2(\mathcal{G}), \tag{14}$$

and dividing by T (the averaging convention of Eq. 3) yields Eq. 4. This route makes clear that the constant 2 is the degree of homogeneity of the variance, and that the result does not depend on the head being linear, only on it being degree-1 homogeneous, which a bias-free ReLU MLP is. It also shows immediately that the identity holds for sum-pooling (degree-1 homogeneous) and, more surprisingly, for max-pooling, since $\max(c\,x) = c\max(x)$ for $c > 0$.

Invariance to the variance convention. Repeating the direct derivation with the biased estimator $\sigma_{\mathrm{b}}^2 = \frac{1}{T}\sum_t (f_t - \mu)^2$ gives $\partial \sigma_{\mathrm{b}}^2 / \partial \tilde{h}_{mk}^{(j)} = \frac{2}{T}(f_j - \mu)\,\partial f_j / \partial \tilde{h}_{mk}^{(j)}$ and therefore $\sum_m E_m^{\sigma^2} = \frac{2}{T^2}\sum_t (f_t - \mu)^2 = \frac{2}{T}\,\sigma_{\mathrm{b}}^2$. The constant is unchanged; the identity holds against *whichever* variance was differentiated. Mixing conventions, backpropagating σ_{b}^2 while normalizing by σ^2, introduces a spurious factor $(T-1)/T$. The normalized shares $\hat{E}_m$ of Eq. 5 are invariant to the choice.

C Proof of Proposition 1

Proof. Let member t have head $\psi_t(z) = w_t^\top z + b_t$, so $\partial f_t / \partial \tilde{h}_{mk}^{(t)} = w_{t,k}/M$ and $p_m^{(t)} := \frac{1}{M}\sum_k w_{t,k}\tilde{h}_{mk}^{(t)}$ satisfies $\sum_m p_m^{(t)} = f_t - b_t$. None of these involve b_t differentiably, so by Lemma 1 the value of $E_m^{\sigma^2}$ depends on the biases only through $f_t - \mu$. Summing over fragments as in Appendix B,

$$\sum_m E_m^{\sigma^2} = \frac{2}{T(T-1)} \sum_t \big(f_t - \mu\big)\big(f_t - b_t\big) \tag{15}$$

$$= \frac{2}{T(T-1)} \left[\sum_t \big(f_t - \mu\big)^2 - \sum_t \big(f_t - \mu\big) b_t \right] \tag{16}$$

$$= \frac{2}{T}\,\sigma^2(\mathcal{G}) \;-\; \frac{2}{T}\,\mathrm{Cov}_t(f_t,\, b_t)\,, \tag{17}$$

using $\sum_t (f_t - \mu) b_t = \sum_t (f_t - \mu)(b_t - \bar{b}) = (T-1)\,\mathrm{Cov}_t(f_t, b_t)$, since $\sum_t (f_t - \mu) = 0$. If the bias is shared, $b_t \equiv b$, the covariance vanishes because $\sum_t (f_t - \mu) = 0$, recovering Eq. 4; the same holds if $b_t \equiv 0$. □

D EXPERIMENTAL DETAILS

D.1 PYTHON ENVIRONMENT.

All experiments were carried out using Python 3.11.5. The principal software dependencies are listed in Table 7; the complete set of requirements is provided in the code repository.

Table 7: **Key libraries used in the experiments.**

Library	Version
PyTorch	2.7.1
PyTorch Geometric	2.5.0
PyTDC	0.4.1
RDKit	2022.9.5
Scikit-learn	1.9.0
Pandas	2.2.1
NumPy	1.26.4

D.2 COMPUTATIONAL RESOURCES.

Experiments were conducted using a single 72-core NVIDIA Grace CPU operating at 3.1 GHz and a single NVIDIA GH200 GPU with 96 GB of memory. Individual runs used up to 32.0 GB of system RAM.

D.3 MOLECULAR GRAPH REPRESENTATION.

Molecules were converted from SMILES strings into undirected molecular graphs using RDKit. Atoms were represented by the 38-dimensional feature vectors listed in Table 8. Bonds determined the graph connectivity.

Table 8: **Atom features used by the GCN models.** Each atom is represented by a 38-dimensional feature vector. Neighbor and hydrogen counts exceeding the listed ranges are mapped to the final category.

Feature	Description	Encoding
Atomic element	Chemical element encoded by atomic number: Be, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, I, or "other"	One-hot (16)
Heavy-neighbor count	Number of directly bonded non-hydrogen atoms	One-hot (6)
Hydrogen count	Total number of attached hydrogens	One-hot (5)
Formal charge	Formal atomic charge	Scalar (1)
Ring membership	Whether the atom belongs to a ring	Binary (1)
Aromaticity	Whether the atom is aromatic	Binary (1)
Explicit valence	Explicit atomic valence	Scalar (1)
Atomic descriptors	Crippen contributions to logP and molar refractivity, TPSA contribution, Labute ASA contribution, EState index, Gasteiger partial charge, and Gasteiger hydrogen partial charge	Continuous (7)

D.4 DATASETS.

We evaluate MUG on six benchmarks from the Therapeutics Data Commons (Huang et al., 2021): **three regression tasks**—Caco-2 permeability (Caco2) (Wang et al., 2016), lipophilicity (Lipophilicity) (Wenlock & Tomkinson, 2015), and aqueous solubility (AqSolDB) (Sorkun et al., 2019)—and **three binary classification tasks**—blood–brain barrier penetration (BBBP) (Martins et al., 2012), Tox21 NR-ER (Huang et al., 2016), and CYP3A4 inhibition (CYP3A4) (Veith et al., 2009).

D.5 MODEL ARCHITECTURE.

Unless otherwise stated, each predictor was a graph convolutional network (GCN). Each graph convolutional layer was followed by dropout and a ReLU activation. The resulting node embeddings after the final ReLU activation were aggregated using global mean pooling, and a linear output layer produced one scalar per molecule.

D.6 TRAINING AND MODEL SELECTION.

Data were scaffold-split into training, validation, and test sets ($70/10/20\%$). Models were trained with AdamW using binary cross-entropy with logits for classification and mean-squared error for regression. A StepLR scheduler halved the learning rate every 30 epochs, and weight decay was fixed at 5×10^{-4}. Models were selected on the validation set using RMSE for regression, AUPRC for BBBP and Tox21, and ROC-AUC for CYP3A4, with early stopping after 15 epochs without improvement.

Hyperparameters were selected independently for each dataset based on validation-set performance, using the search space in Table 9. The selected configurations are reported in Table 10. Predictive performance of the final models is reported in Table 11 for binary classification and Table 12 for regression.

Table 9: **Hyperparameter search space.** For AqSolDB and CYP3A4, batch sizes $\{64, 128\}$ were used.

Hyperparameter	**Candidate values**
Learning rate	$\{1 \times 10^{-3},\ 3 \times 10^{-3},\ 5 \times 10^{-3}\}$
Batch size	$\{32,\ 64\}$
Number of GCN layers	$\{2,\ 3\}$
Hidden dimension	$\{128,\ 256\}$
Dropout probability	$\{0.1,\ 0.2,\ 0.3\}$

Table 10: **Hyperparameters selected for the final models.** LR: learning rate; BS: batch size; L: number of GCN layers; d_h: hidden dimension; p: dropout probability.

Dataset	**LR**	**BS**	$\boldsymbol{L}$	$\boldsymbol{d_h}$	$\boldsymbol{p}$
Caco-2	5×10^{-3}	32	3	128	0.2
Lipo.	3×10^{-3}	32	3	256	0.3
AqSolDB	3×10^{-3}	128	3	256	0.1
BBBP	1×10^{-3}	32	3	256	0.1
Tox21	5×10^{-3}	64	3	256	0.3
CYP3A4	3×10^{-3}	128	3	256	0.1

Table 11: **Predictive performance on the binary-classification datasets.** Values are reported as mean $\pm$ sample standard deviation over five model seeds.

Dataset	Partition	Accuracy ↑	ROC–AUC ↑	AUPRC ↑
BBBP	Train	0.902 ± 0.007	0.959 ± 0.004	0.985 ± 0.002
	Validation	0.891 ± 0.010	0.918 ± 0.003	0.974 ± 0.001
	Test	0.834 ± 0.017	0.882 ± 0.003	0.958 ± 0.002
Tox21 (NR-ER)	Train	0.715 ± 0.029	0.808 ± 0.010	0.474 ± 0.012
	Validation	0.615 ± 0.041	0.693 ± 0.005	0.419 ± 0.006
	Test	0.620 ± 0.037	0.701 ± 0.008	0.413 ± 0.025
CYP3A4	Train	0.848 ± 0.012	0.926 ± 0.009	0.884 ± 0.013
	Validation	0.784 ± 0.007	0.868 ± 0.004	0.834 ± 0.006
	Test	0.792 ± 0.007	0.875 ± 0.006	0.845 ± 0.008

Table 12: Predictive performance on the regression datasets. Values are reported as mean ± sample standard deviation over five model seeds.

Dataset	Partition	RMSE ↓	MAE ↓	R^2 ↑
Caco-2	Train	0.498 ± 0.017	0.402 ± 0.014	0.613 ± 0.027
	Validation	0.518 ± 0.012	0.392 ± 0.014	0.546 ± 0.020
	Test	0.558 ± 0.023	0.460 ± 0.021	0.336 ± 0.054
Lipophilicity	Train	0.675 ± 0.036	0.532 ± 0.032	0.686 ± 0.033
	Validation	0.774 ± 0.016	0.587 ± 0.020	0.566 ± 0.018
	Test	0.771 ± 0.029	0.619 ± 0.031	0.577 ± 0.031
AqSolDB	Train	1.005 ± 0.056	0.703 ± 0.048	0.801 ± 0.022
	Validation	1.169 ± 0.021	0.861 ± 0.014	0.798 ± 0.007
	Test	1.287 ± 0.024	0.962 ± 0.015	0.664 ± 0.013

Table 13: **Influence of the number of Monte Carlo samples on ΔFid.** Values are reported as mean ± sample standard deviation over five model seeds.

	Number of MC samples							
Dataset	30	50	100	400	600	800	1600	2000
AqSolDB	0.253 ± 0.051	0.272 ± 0.054	0.290 ± 0.054	0.299 ± 0.063	0.305 ± 0.059	0.303 ± 0.063	0.308 ± 0.063	0.307 ± 0.063
BBBP	0.150 ± 0.017	0.156 ± 0.015	0.163 ± 0.010	0.170 ± 0.012	0.176 ± 0.017	0.172 ± 0.015	0.176 ± 0.014	0.177 ± 0.011
CYP3A4	0.109 ± 0.024	0.110 ± 0.030	0.114 ± 0.029	0.118 ± 0.031	0.119 ± 0.025	0.118 ± 0.024	0.121 ± 0.026	0.119 ± 0.027
Caco2	0.101 ± 0.025	0.118 ± 0.030	0.149 ± 0.038	0.185 ± 0.047	0.188 ± 0.053	0.196 ± 0.046	0.197 ± 0.049	0.198 ± 0.048
Lipophilicity	0.159 ± 0.019	0.172 ± 0.018	0.178 ± 0.017	0.188 ± 0.015	0.190 ± 0.019	0.189 ± 0.015	0.191 ± 0.017	0.191 ± 0.019
Tox21	0.164 ± 0.037	0.185 ± 0.041	0.199 ± 0.040	0.214 ± 0.047	0.218 ± 0.043	0.221 ± 0.048	0.223 ± 0.042	0.225 ± 0.047

D.7 Number of Monte Carlo Dropout Samples

Table 13 examines the sensitivity of $\Delta\mathrm{Fid}$ to the number of Monte Carlo dropout samples. Fidelity generally improves as the number of samples increases, with the largest gains occurring between 30 and 400 samples. Beyond approximately 800 samples, the results stabilize: increasing the sample count from 800 to 2000 changes the macro-average $\Delta\mathrm{Fid}$ only from 0.200 to 0.203, and the differences for individual datasets are small relative to the variation across trained models. We therefore use $T = 800$ throughout the main experiments, as it provides a reliable estimate of predictive uncertainty while avoiding the substantially greater computational cost of 1600 or 2000 stochastic forward passes.

E EVALUATION METRICS

Let $\sigma(\mathcal{G})$ denote the predictive standard deviation for molecule $\mathcal{G}$. For occlusion metrics, we rank atoms by signed attribution and set $k = \lceil 0.1M \rceil$, where M is the number of atoms. The sets top_k and bot_k contain the k highest- and lowest-ranked atoms, respectively. We write $\mathcal{G} \setminus S$ for the graph with the input features of atoms in S zeroed, leaving topology unchanged.

(i) Relative completeness residual. We measure the relative discrepancy between summed atom-level attributions and the scaled predictive variance specified by Theorem 1:

$$r(\mathcal{G}) = \frac{\left| \sum_m E_m^{\sigma^2} - \frac{2}{T}\sigma^2(\mathcal{G}) \right|}{\frac{2}{T}\sigma^2(\mathcal{G})}. \tag{18}$$

Lower values indicate better completeness. For MUG, the residual is zero up to numerical precision when the theorem's assumptions hold. The relative residual is defined for $\sigma^2(\mathcal{G}) > 0$.

(ii) Top-ranked occlusion fidelity. We measure the relative reduction in predictive standard deviation after masking the highest-ranked atoms:

$$\mathrm{Fid}^{\downarrow}(\mathcal{G}, k) = \frac{\sigma(\mathcal{G}) - \sigma(\mathcal{G} \setminus \mathrm{top}_k)}{\sigma(\mathcal{G})}. \tag{19}$$

Positive values indicate reduced uncertainty; negative values indicate increased uncertainty.

(iii) Controlled fidelity. To control for the number of masked atoms, we compare top-ranked occlusion with masking the same number of lowest-ranked atoms:

$$\begin{aligned} \mathrm{Fid}^{\uparrow}(\mathcal{G}, k) &= \frac{\sigma(\mathcal{G}) - \sigma(\mathcal{G} \setminus \mathrm{bot}_k)}{\sigma(\mathcal{G})}, \\ \Delta\mathrm{Fid}(\mathcal{G}, k) &= \mathrm{Fid}^{\downarrow}(\mathcal{G}, k) - \mathrm{Fid}^{\uparrow}(\mathcal{G}, k). \end{aligned} \tag{20}$$

Positive controlled fidelity indicates that top-ranked masking yields lower uncertainty than bottom-ranked masking.

(iv) Selectivity rate. To complement the magnitude-sensitive fidelity scores, we measure how often top-ranked masking yields lower uncertainty than bottom-ranked masking, assigning half credit to ties. For N molecules with valid scores, let $\Delta_i = \Delta\mathrm{Fid}(\mathcal{G}_i, k_i)$. We define

$$\mathrm{SR} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{1}[\Delta_i > 0] + \frac{1}{2}\mathbf{1}[\Delta_i = 0] \right). \tag{21}$$

Unlike mean controlled fidelity, selectivity depends only on the direction of the comparison, preventing large normalized scores from dominating the result.

Table 14: **Aggregate completeness residuals.** Completeness residual $r(\mathcal{G})$ (Eq. 18) for MUG and alternative attribution variants. We compute the median and P95 separately within each dataset and report their equally weighted means across the six datasets. Results use a GCN backbone, a scaffold split, five seeds, and $T = 800$ MC-dropout samples. †Input saliency is included as a non-decomposition negative control.

Attribution	Mean of dataset medians ↓	Mean of dataset P95 values ↓
MUG	$\mathbf{(7.98 \pm 0.27) \times 10^{-6}}$	$\mathbf{(3.69 \pm 0.25) \times 10^{-5}}$
$\mathrm{ReLU(MUG)}$	$(8.10 \pm 0.27) \times 10^{-6}$	$(5.71 \pm 2.57) \times 10^{-5}$
$\lvert\mathrm{MUG}\rvert$	$(8.10 \pm 0.27) \times 10^{-6}$	$(7.25 \pm 5.06) \times 10^{-5}$
Input saliency$(\sigma^2)^\dagger$	$(2.29 \pm 0.09) \times 10^{3}$	$(4.21 \pm 0.23) \times 10^{3}$

Table 15: **Per-dataset completeness residuals.** For each dataset and seed, we compute the median and P95 of the molecule-level completeness residuals. Values are the mean ± standard deviation across five random seeds. Results use a GCN backbone, a scaffold split, and $T = 800$ MC-dropout samples. †Input saliency is included as a non-decomposition negative control.

Attribution	Statistic	Caco2	Lipophilicity	AqSolDB	BBBP	Tox21	CYP3A4
MUG	Median	$\mathbf{(3.50 \pm 0.12) \times 10^{-5}}$	$\mathbf{(3.46 \pm 0.26) \times 10^{-6}}$	$\mathbf{(7.10 \pm 0.71) \times 10^{-6}}$	$\mathbf{(1.68 \pm 0.34) \times 10^{-6}}$	$\mathbf{(2.01 \pm 0.21) \times 10^{-7}}$	$\mathbf{(4.31 \pm 0.64) \times 10^{-7}}$
	P95	$\mathbf{(1.20 \pm 0.13) \times 10^{-4}}$	$\mathbf{(2.61 \pm 0.33) \times 10^{-5}}$	$\mathbf{(4.86 \pm 0.38) \times 10^{-5}}$	$\mathbf{(1.77 \pm 0.19) \times 10^{-5}}$	$\mathbf{(3.36 \pm 0.42) \times 10^{-6}}$	$\mathbf{(6.11 \pm 0.77) \times 10^{-6}}$
$\mathrm{ReLU(MUG)}$	Median	$(3.50 \pm 0.12) \times 10^{-5}$	$(3.55 \pm 0.24) \times 10^{-6}$	$(7.64 \pm 0.55) \times 10^{-6}$	$(1.70 \pm 0.35) \times 10^{-6}$	$(2.18 \pm 0.21) \times 10^{-7}$	$(4.93 \pm 1.13) \times 10^{-7}$
	P95	$(1.20 \pm 0.13) \times 10^{-4}$	$(2.80 \pm 0.33) \times 10^{-5}$	$(7.30 \pm 2.11) \times 10^{-5}$	$(1.88 \pm 0.15) \times 10^{-5}$	$(7.95 \pm 6.28) \times 10^{-6}$	$(9.51 \pm 15.13) \times 10^{-5}$
$\lvert\mathrm{MUG}\rvert$	Median	$(3.50 \pm 0.12) \times 10^{-5}$	$(3.55 \pm 0.24) \times 10^{-6}$	$(7.64 \pm 0.55) \times 10^{-6}$	$(1.70 \pm 0.35) \times 10^{-6}$	$(2.18 \pm 0.21) \times 10^{-7}$	$(4.93 \pm 1.13) \times 10^{-7}$
	P95	$(1.20 \pm 0.13) \times 10^{-4}$	$(2.82 \pm 0.32) \times 10^{-5}$	$(7.64 \pm 2.49) \times 10^{-5}$	$(1.88 \pm 0.15) \times 10^{-5}$	$(8.59 \pm 7.63) \times 10^{-6}$	$(1.83 \pm 3.04) \times 10^{-4}$
Input saliency$(\sigma^2)^\dagger$	Median	$(1.63 \pm 0.22) \times 10^{3}$	$(2.13 \pm 0.13) \times 10^{3}$	$(2.73 \pm 0.21) \times 10^{3}$	$(1.37 \pm 0.11) \times 10^{3}$	$(2.67 \pm 0.30) \times 10^{3}$	$(3.18 \pm 0.20) \times 10^{3}$
	P95	$(1.83 \pm 0.21) \times 10^{3}$	$(4.24 \pm 0.51) \times 10^{3}$	$(4.52 \pm 0.33) \times 10^{3}$	$(3.60 \pm 0.92) \times 10^{3}$	$(5.45 \pm 0.39) \times 10^{3}$	$(5.60 \pm 0.51) \times 10^{3}$

F Additional Results

F.1 MUG exact variance decomposition

Tables 14 and 15 report the empirical completeness residual defined in Eq. 18. MUG yields the lowest aggregate residuals, with a mean dataset median of 7.98×10^{-6} and a mean dataset P95 of 3.69×10^{-5}, and maintains near-zero residuals across all six datasets. Applying ReLU or taking absolute values produces similar median residuals but consistently larger P95 values, indicating that discarding signed contributions primarily degrades completeness in the tails of the molecule-level distribution. In contrast, input saliency yields residuals on the order of 10^3—approximately eight orders of magnitude larger than those of MUG—confirming that it does not provide an additive decomposition of predictive variance.

F.2 Computational cost

Figure 6 summarizes the offline training and online attribution costs.

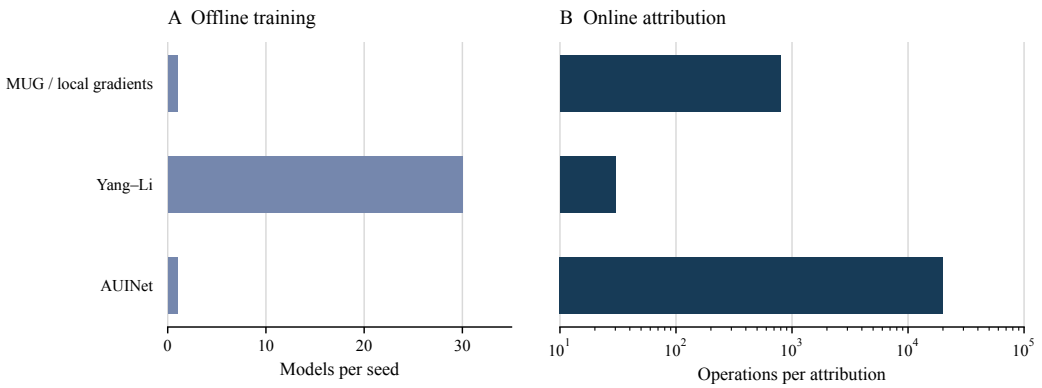


Figure 6: **Offline and online computational requirements.** The left panel reports the number of independently trained models per dataset and seed. The right panel reports method-specific operations for one atom attribution on a logarithmic scale. For the MUG / local gradients methods, values count MC forward–backward sample equivalents with $T = 800$; Yang–Li uses $E = 30$ ensemble-member forward passes. AUINet averages 10 uncertainty estimates, each comprising 10 refinement iterations with 200 MC-dropout samples, resulting in $10 \times 10 \times 200 = 20,000$ internal sample updates.

F.3 What the model has not seen: localized uncertainty around unseen atom types

Figure 7 shows the distribution of uncertainty assigned to entire molecules under three training scenarios. The reference model is trained on the complete CYP3A4 training set. The F-excluded model is trained after removing all fluorine-containing molecules, whereas the random-removal control is trained after removing the same number of randomly selected molecules. The reference and random-removal models exhibit similar uncertainty distributions. In contrast, the F-excluded model shows a large distributional shift for fluorinated molecules, with the magnitude of the shift increasing with the number of fluorine atoms in a molecule.

We next test whether the same behavior extends from an unseen atom type to an unseen molecular motif. We exclude sulfone-containing molecules from the Lipophilicity training set and compare the resulting model with the reference and random-removal models. Figure 8 shows that the motif-excluded model assigns a larger fraction of its MUG attribution directly to sulfone-group atoms and their immediate neighborhoods. The separation is greatest at short topological distances and decreases as the radius expands, demonstrating that the additional uncertainty is localized around the withheld motif.

At the molecule level, sulfone exclusion produces a smaller shift than fluorine exclusion. Sulfone-containing test molecules move modestly from the lowest normalized uncertainty interval toward the next interval (Fig. 9). We hypothesize that excluding a molecular motif creates a less unfamiliar chemical environment than excluding an entire atom type. In the atom-exclusion setting, the model encounters an atom type that was absent from the training data. In the motif-exclusion setting, the individual atom types are known to the model, and only their specific arrangement and bonding topology are unseen.

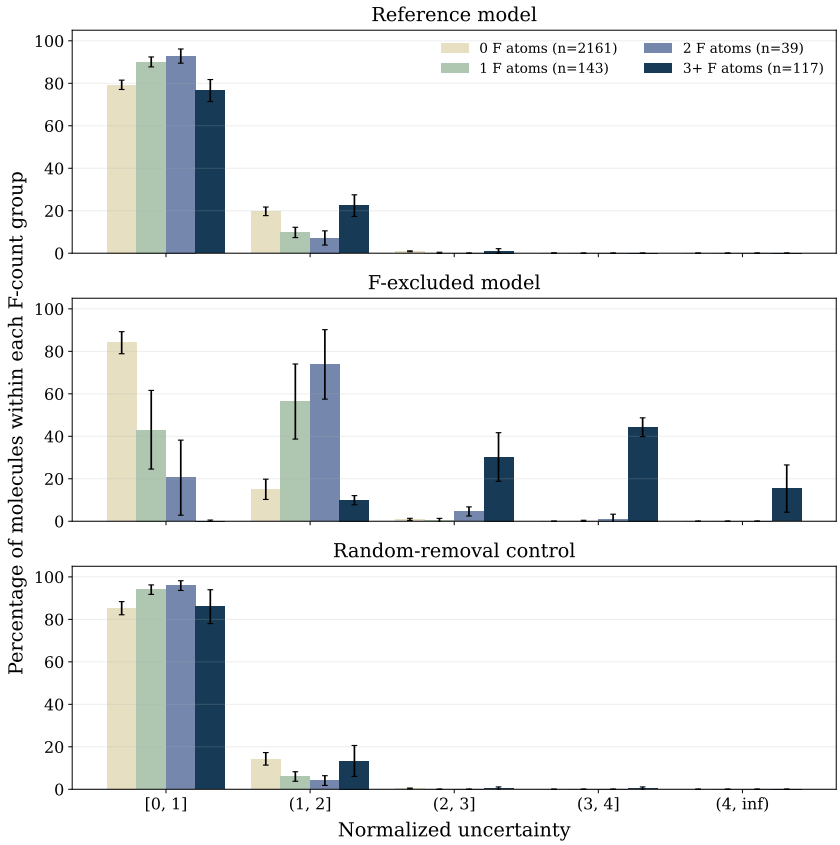


Figure 7: **Excluding fluorine-containing molecules from training selectively increases uncertainty for fluorinated test molecules.** Distribution of normalized epistemic uncertainty in the CYP3A4 test set, stratified by the number of fluorine atoms in each molecule. Predictive standard deviations are normalized by the pooled 80th percentile obtained from the reference models. Bars show the mean percentage of molecules in each uncertainty interval, and error bars denote $\pm$ one standard deviation across five random seeds. Under fluorine exclusion, fluorinated molecules shift toward higher uncertainty intervals, with the strongest shift for molecules containing three or more fluorine atoms. This pattern is absent in the reference model and the random-removal control, indicating that it results from the systematic exclusion of fluorine-containing training molecules rather than from the reduced training-set size. Group sizes are reported in the legend.

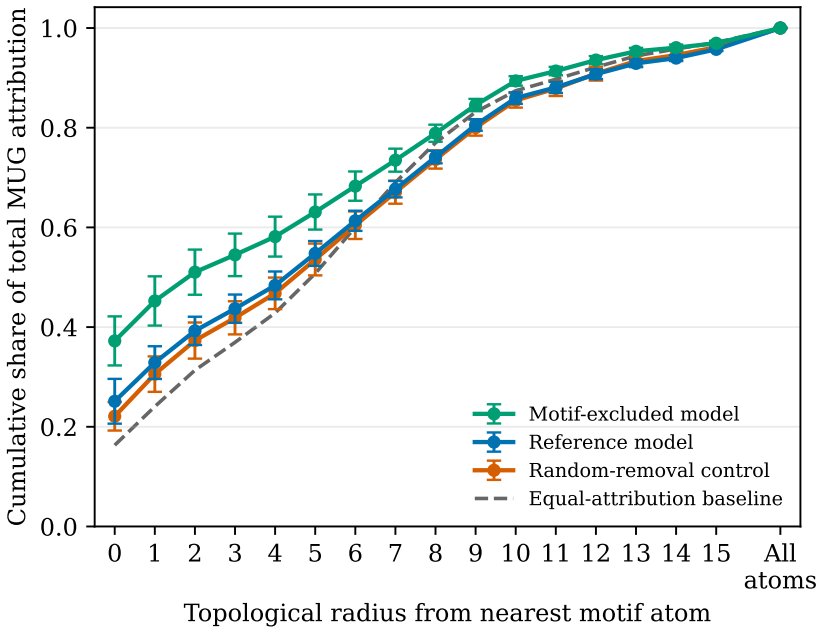


Figure 8: **Models trained without sulfone-containing molecules concentrate uncertainty around sulfone groups in the test set.** This result is shown for the Lipophilicity dataset. Curves represent cumulative MUG attribution within r bonds of the nearest sulfone-group atom.

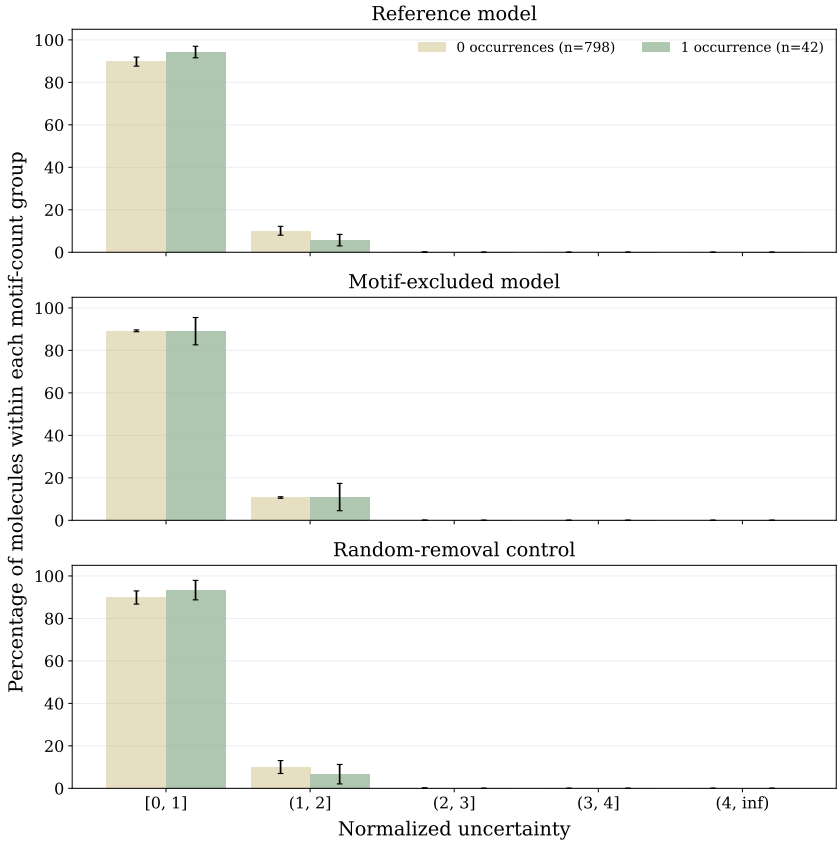


Figure 9: **Excluding sulfone-containing molecules from training selectively increases uncertainty for sulfone-containing test molecules.** Distribution of normalized epistemic uncertainty in the Lipophilicity test set. Predictive standard deviations are normalized by the pooled 80th percentile obtained from the reference models. Bars show the mean percentage of molecules in each uncertainty interval, and error bars denote ± one standard deviation across five random seeds.

Table 16: **All occlusion-fidelity metrics on Caco2.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	ΔFid ↑	Fid$^{\downarrow}$ ↑	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	0.003 ± 0.004	0.047 ± 0.017	78.1 ± 8.8	22.4 ± 7.7	52.1 ± 1.7	15.4 ± 2.3	0.042 ± 0.034
Degree	0.165 ± 0.073	0.096 ± 0.054	87.0 ± 17.0	$\underline{85.9 \pm 6.7}$	91.8 ± 10.3	$\underline{75.2 \pm 20.1}$	0.835 ± 0.206
Saliency	$\mathbf{0.212 \pm 0.067}$	$\mathbf{0.134 \pm 0.044}$	92.6 ± 10.9	$\mathbf{87.7 \pm 5.1}$	$\mathbf{96.0 \pm 5.9}$	$\mathbf{81.0 \pm 14.4}$	$\mathbf{0.921 \pm 0.119}$
GNNExplainer	0.016 ± 0.009	0.051 ± 0.016	80.7 ± 7.9	27.3 ± 7.7	58.0 ± 5.1	20.8 ± 3.5	0.160 ± 0.102
Yang & Li (aleatoric)	0.097 ± 0.018	0.073 ± 0.017	85.5 ± 11.6	78.1 ± 12.1	89.9 ± 8.7	65.1 ± 14.5	0.798 ± 0.174
Yang & Li (epistemic)	-0.134 ± 0.111	-0.268 ± 0.123	6.6 ± 5.2	69.2 ± 5.3	25.9 ± 12.5	4.3 ± 3.0	-0.481 ± 0.250
Yang & Li (total)	0.014 ± 0.037	-0.015 ± 0.034	45.1 ± 18.4	71.3 ± 12.8	61.0 ± 16.4	37.6 ± 18.6	0.220 ± 0.328
AUINet	0.036 ± 0.011	0.054 ± 0.015	$\mathbf{99.7 \pm 0.7}$	23.1 ± 15.5	82.6 ± 10.7	22.9 ± 15.4	0.653 ± 0.215
MUG	$\underline{0.194 \pm 0.051}$	$\underline{0.132 \pm 0.040}$	$\underline{92.7 \pm 9.5}$	76.8 ± 2.3	$\underline{94.2 \pm 7.5}$	71.3 ± 8.8	$\underline{0.884 \pm 0.151}$

F.4 Occlusion Fidelity of Uncertainty Attributions

Evaluation protocol. We provide implementation details for the comparisons in §4.2. For each molecule with M atoms, we rank the atom-level scores on the unmodified graph and separately mask the $k = \lceil 0.1M \rceil$ highest- and lowest-ranked atoms. Scores are ranked by their signed values, without taking absolute values. Masking sets the selected input atom-feature vectors to zero while retaining all atoms, bonds, and bond-only features. For the bond-message representations used by Yang & Li and AUINet, we also zero the corresponding source-atom features embedded in outgoing directed-bond feature vectors. We then recompute uncertainty on each masked graph without retraining the model.

Yang & Li. We use the authors' released implementation of Yang & Li (2023). Each ensemble comprises $E = 30$ independently trained members optimized using the released heteroscedastic Gaussian loss. After training, we apply the authors' post-hoc recalibration procedure, jointly fine-tuning the variance heads while keeping all remaining model parameters fixed. Evaluation is restricted to the three regression datasets, Caco2, Lipophilicity, and AqSolDB, because the released implementation supports only regression. The Yang & Li row in Table 1 reports the *aleatoric* component: atoms are ranked by their predicted atomic variances averaged over ensemble members. The implementation also evaluates epistemic uncertainty, defined as the variance across ensemble-member predictions, and total uncertainty, defined as the sum of the aleatoric and epistemic variances. For each component, the atom ranking is paired with the corresponding molecular uncertainty component.

AUINet. We use the released AUINet implementation of Qian et al. (2026). Classification uses binary cross-entropy, whereas our regression adapter uses a Gaussian negative log-likelihood with a positive scale obtained through a softplus transformation. We use ten refinement iterations with 200 internal MC-dropout samples per iteration and average $R = 10$ repeated uncertainty estimates. Atom-level uncertainty is computed from the representation variances returned at the final refinement iteration: we take the square root of each coordinate-wise variance and then average over hidden coordinates and repetitions. Following the molecule-level aggregation described in the AUINet paper, molecular uncertainty is defined as the mean of these atom-level scores.

Detailed results. Detailed occlusion results are reported separately for Caco2 (Table 16), Lipophilicity (Table 17), AqSolDB (Table 18), BBBP (Table 19), Tox21 (Table 20), and CYP3A4 (Table 21). These tables extend the main comparison by reporting both the magnitude of uncertainty changes and their directional consistency across test molecules.

Table 17: **All occlusion-fidelity metrics on Lipophilicity.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	0.001 ± 0.006	-0.010 ± 0.017	53.5 ± 6.1	46.2 ± 6.6	49.1 ± 0.7	22.9 ± 1.1	-0.018 ± 0.014
Degree	0.125 ± 0.150	-0.050 ± 0.052	41.4 ± 6.7	58.0 ± 11.1	52.5 ± 10.1	24.6 ± 6.9	0.049 ± 0.203
Saliency	0.176 ± 0.010	0.125 ± 0.022	79.7 ± 4.7	71.8 ± 5.0	83.6 ± 3.1	56.6 ± 1.8	0.672 ± 0.061
GNNExplainer	0.007 ± 0.005	-0.006 ± 0.018	55.0 ± 6.1	47.2 ± 6.9	52.1 ± 1.1	24.0 ± 2.2	0.042 ± 0.021
Yang & Li (aleatoric)	0.133 ± 0.008	0.120 ± 0.008	92.0 ± 1.7	61.6 ± 7.4	$\mathbf{93.4 \pm 0.7}$	54.9 ± 6.0	$\mathbf{0.868 \pm 0.015}$
Yang & Li (epistemic)	-0.024 ± 0.081	-0.204 ± 0.074	21.0 ± 4.7	$\mathbf{88.1 \pm 2.7}$	48.6 ± 10.5	17.2 ± 3.8	-0.029 ± 0.210
Yang & Li (total)	0.066 ± 0.016	0.037 ± 0.018	62.9 ± 6.9	69.3 ± 3.0	71.1 ± 6.0	41.5 ± 4.0	0.422 ± 0.121
AUINet	-0.008 ± 0.005	0.029 ± 0.007	$\mathbf{92.6 \pm 5.3}$	0.9 ± 0.9	40.3 ± 5.7	0.8 ± 0.8	-0.195 ± 0.115
MUG	$\mathbf{0.190 \pm 0.017}$	$\mathbf{0.153 \pm 0.025}$	84.4 ± 3.7	70.3 ± 9.3	87.7 ± 1.6	$\mathbf{59.1 \pm 5.6}$	0.754 ± 0.033

Table 18: **All occlusion-fidelity metrics on AqSolDB.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	-0.008 ± 0.009	-0.037 ± 0.050	50.6 ± 5.2	48.0 ± 4.4	49.4 ± 1.2	24.0 ± 1.0	-0.012 ± 0.024
Degree	0.261 ± 0.202	0.001 ± 0.193	69.9 ± 11.2	88.7 ± 2.8	79.4 ± 10.9	65.0 ± 9.8	0.589 ± 0.218
Saliency	0.296 ± 0.035	$\mathbf{0.151 \pm 0.039}$	$\mathbf{86.1 \pm 4.4}$	83.1 ± 2.9	$\mathbf{89.5 \pm 4.0}$	72.1 ± 3.7	$\mathbf{0.791 \pm 0.079}$
GNNExplainer	0.026 ± 0.010	-0.022 ± 0.052	55.2 ± 5.5	52.6 ± 4.6	55.2 ± 1.8	28.4 ± 0.9	0.103 ± 0.037
Yang & Li (aleatoric)	0.025 ± 0.043	-0.178 ± 0.112	18.4 ± 13.0	$\mathbf{96.2 \pm 2.3}$	57.4 ± 12.5	16.9 ± 11.7	0.147 ± 0.251
Yang & Li (epistemic)	-0.043 ± 0.056	-0.092 ± 0.070	32.7 ± 14.4	61.6 ± 10.7	42.1 ± 12.1	14.8 ± 5.8	-0.159 ± 0.242
Yang & Li (total)	-0.015 ± 0.043	-0.185 ± 0.100	14.9 ± 10.0	91.6 ± 4.2	47.1 ± 12.1	12.0 ± 7.5	-0.057 ± 0.241
AUINet	-0.028 ± 0.018	0.009 ± 0.015	61.1 ± 12.1	8.1 ± 2.3	37.3 ± 12.1	6.1 ± 2.5	-0.255 ± 0.243
MUG	$\mathbf{0.304 \pm 0.061}$	0.151 ± 0.064	84.5 ± 6.0	89.6 ± 1.9	88.8 ± 6.1	$\mathbf{76.0 \pm 5.2}$	0.776 ± 0.121

Table 19: **All occlusion-fidelity metrics on BBBP.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	0.005 ± 0.013	-0.003 ± 0.029	62.2 ± 5.6	38.3 ± 3.7	51.8 ± 1.9	19.9 ± 2.2	0.035 ± 0.038
Degree	0.082 ± 0.048	0.019 ± 0.050	67.7 ± 7.3	$\mathbf{57.2 \pm 2.5}$	65.6 ± 3.3	38.5 ± 2.4	0.311 ± 0.066
Saliency	$\mathbf{0.179 \pm 0.015}$	0.108 ± 0.016	84.1 ± 3.8	$\mathbf{57.2 \pm 5.5}$	84.7 ± 1.8	$\mathbf{48.3 \pm 3.7}$	0.695 ± 0.036
GNNExplainer	0.023 ± 0.011	0.009 ± 0.026	65.1 ± 6.2	41.3 ± 5.5	52.9 ± 3.2	24.4 ± 2.4	0.057 ± 0.064
Yang & Li (aleatoric)	–	–	–	–	–	–	–
Yang & Li (epistemic)	–	–	–	–	–	–	–
Yang & Li (total)	–	–	–	–	–	–	–
AUINet	0.035 ± 0.006	0.070 ± 0.008	$\mathbf{100.0 \pm 0.0}$	0.0 ± 0.0	$\mathbf{92.6 \pm 1.7}$	0.0 ± 0.0	$\mathbf{0.853 \pm 0.034}$
MUG	0.175 ± 0.013	$\mathbf{0.121 \pm 0.027}$	82.4 ± 4.3	53.9 ± 9.1	82.8 ± 2.1	44.5 ± 6.4	0.655 ± 0.041

Table 20: **All occlusion-fidelity metrics on Tox21.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	ΔFid ↑	$\text{Fid}^{\downarrow}$ ↑	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	0.006 ± 0.003	-0.025 ± 0.019	53.7 ± 3.2	47.0 ± 2.7	50.9 ± 1.0	22.7 ± 1.2	0.017 ± 0.020
Degree	$\mathbf{0.332 \pm 0.110}$	0.092 ± 0.021	67.3 ± 3.5	69.7 ± 7.9	71.8 ± 6.1	49.6 ± 7.3	0.436 ± 0.121
Saliency	0.216 ± 0.046	$\mathbf{0.133 \pm 0.015}$	$\mathbf{80.4 \pm 2.6}$	70.1 ± 6.6	84.3 ± 2.7	56.2 ± 4.9	0.686 ± 0.054
GNNExplainer	0.014 ± 0.020	-0.019 ± 0.004	54.2 ± 1.7	48.4 ± 2.6	51.6 ± 2.9	24.1 ± 1.0	0.033 ± 0.057
Yang & Li (aleatoric)	–	–	–	–	–	–	–
Yang & Li (epistemic)	–	–	–	–	–	–	–
Yang & Li (total)	–	–	–	–	–	–	–
AUINet	-0.030 ± 0.026	0.017 ± 0.021	65.0 ± 27.9	1.8 ± 1.4	28.6 ± 17.6	1.3 ± 1.0	-0.428 ± 0.351
MUG	0.221 ± 0.043	0.120 ± 0.025	79.9 ± 3.0	$\mathbf{79.6 \pm 3.8}$	$\mathbf{84.8 \pm 2.5}$	$\mathbf{63.7 \pm 4.0}$	$\mathbf{0.697 \pm 0.050}$

Table 21: **All occlusion-fidelity metrics on CYP3A4.** Cells report mean ± sample standard deviation over five seeds. Best five-seed means are bold and second-best are underlined.

Atom attribution	$\Delta\mathrm{Fid}\uparrow$	$\mathrm{Fid}^{\downarrow}\uparrow$	Top decrease rate (%) ↑	Bottom non-decrease rate (%) ↑	Selectivity rate (%) ↑	Strong directional rate (%) ↑	Mean direction ↑
Random	0.001 ± 0.007	-0.070 ± 0.028	43.8 ± 2.9	56.5 ± 3.1	49.5 ± 0.8	22.5 ± 1.1	-0.011 ± 0.017
Degree	-0.241 ± 0.103	-0.346 ± 0.109	22.5 ± 7.5	71.4 ± 5.1	34.5 ± 8.1	12.9 ± 4.3	-0.309 ± 0.162
Saliency	$\underline{0.110 \pm 0.019}$	$\mathbf{0.068 \pm 0.025}$	$\mathbf{76.1 \pm 3.7}$	$\underline{76.7 \pm 5.6}$	$\mathbf{80.8 \pm 2.1}$	$\underline{57.5 \pm 3.0}$	$\mathbf{0.617 \pm 0.043}$
GNNExplainer	0.013 ± 0.007	-0.064 ± 0.029	45.0 ± 3.5	57.7 ± 3.3	51.7 ± 1.1	24.0 ± 0.6	0.034 ± 0.022
Yang & Li (aleatoric)	–	–	–	–	–	–	–
Yang & Li (epistemic)	–	–	–	–	–	–	–
Yang & Li (total)	–	–	–	–	–	–	–
AUINet	-0.023 ± 0.011	0.012 ± 0.009	66.8 ± 14.7	1.4 ± 2.1	27.2 ± 13.6	1.0 ± 1.6	-0.456 ± 0.272
MUG	$\mathbf{0.119 \pm 0.027}$	$\underline{0.063 \pm 0.036}$	$\underline{75.7 \pm 2.9}$	$\mathbf{83.6 \pm 3.9}$	$\underline{79.1 \pm 2.7}$	$\mathbf{63.0 \pm 2.6}$	$\underline{0.583 \pm 0.055}$

F.5 Details of Prediction and Uncertainty Attribution

F.5.1 Attribution comparison and fidelity metrics

We compare prediction and uncertainty attribution separately for each molecule. Let $\mathbf{a}_i^{\mathrm{Pred}}$ and $\mathbf{a}_i^{\mathrm{MUG}}$ denote the atom-level prediction and uncertainty attribution vectors for molecule i. Their agreement is measured using the Spearman correlation between absolute attribution values:

$$\rho_i = \rho_{\mathrm{Spearman}} \left(\left| \mathbf{a}_i^{\mathrm{Pred}} \right|, \left| \mathbf{a}_i^{\mathrm{MUG}} \right| \right). \tag{22}$$

Absolute values are used because attribution signs refer to different model outputs and are therefore not directly comparable. We additionally measure whether the most important atom sets of the two methods are disjoint:

$$D_i = \mathbb{I} \left[T_i^{\mathrm{Pred}} \cap T_i^{\mathrm{MUG}} = \varnothing \right], \tag{23}$$

where T_i^{Pred} and T_i^{MUG} contain the atoms with the largest absolute attribution, including ties. Thus, $D_i = 1$ indicates that prediction and uncertainty attribution select different top atoms.

For prediction fidelity, atoms are ranked by absolute attribution. We mask the $k = \lceil 0.1 M_i \rceil$ most and least important atoms separately and measure the corresponding absolute changes in the predictive mean:

$$D_i^{\mathrm{top}} = \left| \mu_i - \mu_i^{\mathrm{top}} \right|, \qquad D_i^{\mathrm{bottom}} = \left| \mu_i - \mu_i^{\mathrm{bottom}} \right|. \tag{24}$$

Prediction fidelity is aggregated within each dataset d and seed s using a ratio of sums:

$$\Delta\mathrm{Fid}_{d,s}^{\mathrm{Pred}} = \frac{\sum_{i \in (d,s)} \left(D_i^{\mathrm{top}} - D_i^{\mathrm{bottom}} \right)}{\sum_{i \in (d,s)} \left(|\mu_i| + \varepsilon \right)}. \tag{25}$$

A positive value indicates that masking highly attributed atoms changes the predictive mean more strongly than masking atoms with low attribution. Unlike uncertainty fidelity, prediction fidelity is aggregated using a ratio of sums. This prevents molecules with near-zero predictive means from producing disproportionately large normalized scores and dominating the dataset-level result. The same instability was substantially less pronounced for uncertainty fidelity because the predictive standard deviation is non-negative and near-zero denominators had little influence on the aggregate results.

F.5.2 Dataset-level results

Tables 22–24 report results separately for each dataset. Dataset-level values are means $\pm$ standard deviations over five seeds. The macro-average is computed by first averaging the six datasets within each seed and then reporting the mean and standard deviation over seeds.

Table 22: **Dataset-level agreement between prediction and uncertainty attribution.** Each Grad×Input variant is compared with final-layer MUG(σ^2). Spearman correlation is calculated between absolute atom-importance rankings. Top disagreement is the fraction of molecules for which the highest-importance atom sets are disjoint. Dataset-level values are mean ± standard deviation over five seeds.

	Grad×Input, final layer		Grad×Input, input layer	
Dataset	Mean $\rho \downarrow$	Disagreement ↑	Mean $\rho \downarrow$	Disagreement ↑
AqSolDB	0.614 ± 0.019	$(71.7 \pm 2.9)\%$	0.381 ± 0.020	$(81.0 \pm 3.4)\%$
BBBP	0.190 ± 0.047	$(72.6 \pm 2.4)\%$	0.268 ± 0.022	$(78.2 \pm 1.4)\%$
CYP3A4	0.506 ± 0.030	$(73.7 \pm 3.2)\%$	0.371 ± 0.029	$(79.9 \pm 1.9)\%$
Caco2	0.869 ± 0.021	$(56.7 \pm 2.4)\%$	0.630 ± 0.051	$(82.3 \pm 3.6)\%$
Lipophilicity	0.038 ± 0.127	$(71.5 \pm 2.8)\%$	0.337 ± 0.037	$(75.2 \pm 4.4)\%$
Tox21	0.201 ± 0.107	$(78.8 \pm 5.4)\%$	0.271 ± 0.024	$(84.8 \pm 2.1)\%$
Macro-average	$\mathbf{0.403 \pm 0.024}$	$\mathbf{(70.8 \pm 0.6)\%}$	$\mathbf{0.376 \pm 0.012}$	$\mathbf{(80.2 \pm 1.2)\%}$

Table 23: **Dataset-level prediction fidelity.** Atoms are ranked by absolute attribution. Positive values indicate that masking highly attributed atoms changes the predictive mean more strongly than masking atoms with low attribution. Dataset-level values are reported as mean ± standard deviation over five seeds.

Dataset	Grad×Input(μ), final	Grad×Input(μ), input	MUG(σ^2), final
AqSolDB	$+0.183 \pm 0.032$	$+0.404 \pm 0.036$	$+0.114 \pm 0.040$
BBBP	$+0.367 \pm 0.047$	$+0.535 \pm 0.042$	$+0.228 \pm 0.027$
CYP3A4	$+0.487 \pm 0.039$	$+0.634 \pm 0.094$	$+0.408 \pm 0.112$
Caco2	$+0.109 \pm 0.015$	$+0.137 \pm 0.019$	$+0.100 \pm 0.012$
Lipophilicity	$+0.213 \pm 0.007$	$+0.302 \pm 0.028$	$+0.215 \pm 0.021$
Tox21	$+0.198 \pm 0.039$	$+0.486 \pm 0.096$	$+0.190 \pm 0.023$
Macro-average	$+0.259 \pm 0.005$	$\mathbf{+0.417 \pm 0.023}$	$+0.209 \pm 0.025$

Table 24: **Dataset-level uncertainty fidelity.** Atoms are ranked using signed attribution scores. For each molecule, $\Delta\mathrm{Fid}_\sigma$ compares the predictive standard deviation after masking the lowest- and highest-ranked atoms, normalized by the baseline predictive standard deviation. Positive values indicate that masking atoms with high uncertainty attribution reduces predictive uncertainty more strongly than masking bottom-ranked atoms. Dataset-level values are reported as mean ± standard deviation over five seeds.

Dataset	Grad×Input(μ), final	Grad×Input(μ), input	MUG(σ^2), final
AqSolDB	-0.285 ± 0.026	-0.386 ± 0.038	$+0.306 \pm 0.059$
BBBP	-0.040 ± 0.038	$+0.034 \pm 0.048$	$+0.173 \pm 0.013$
CYP3A4	$+0.059 \pm 0.024$	$+0.431 \pm 0.164$	$+0.115 \pm 0.027$
Caco2	-0.203 ± 0.062	-0.252 ± 0.045	$+0.192 \pm 0.048$
Lipophilicity	-0.091 ± 0.024	-0.241 ± 0.039	$+0.188 \pm 0.020$
Tox21	-0.042 ± 0.031	-0.040 ± 0.063	$+0.222 \pm 0.043$
Macro-average	-0.100 ± 0.015	-0.075 ± 0.029	$\mathbf{+0.199 \pm 0.026}$

Table 25: **Unseen molecular fragments receive larger MUG attributions.** Test molecules are decomposed into BRICS fragments and classified as seen when the exact canonical fragment occurs in the training set and unseen otherwise. Atom counts are medians across fragment occurrences. Attributions are reported as mean ± std across five seeds.

Dataset	Group	Median atoms	Mean MUG ($\times 10^{-4}$)	Mean MUG / atom ($\times 10^{-5}$)
BBBP	Seen	3	0.697 ± 0.112	1.789 ± 0.324
	Unseen	10	4.594 ± 0.901	8.213 ± 1.643
CYP3A4	Seen	3	0.986 ± 0.242	2.475 ± 0.582
	Unseen	13	3.948 ± 0.968	3.661 ± 0.953
Caco2	Seen	3	0.104 ± 0.004	0.229 ± 0.007
	Unseen	15	0.453 ± 0.022	0.298 ± 0.015
Lipophilicity	Seen	4	0.192 ± 0.018	0.490 ± 0.057
	Unseen	11	0.711 ± 0.060	0.717 ± 0.065
AqSolDB	Seen	3	0.169 ± 0.015	0.431 ± 0.026
	Unseen	12	0.964 ± 0.140	0.709 ± 0.098
Tox21	Seen	3	0.474 ± 0.071	1.074 ± 0.158
	Unseen	11	2.777 ± 0.518	2.603 ± 0.538

Table 26: **Unseen molecular fragments receive larger MUG attributions.** BRICS fragment statistics for exact seen, similar seen, and unseen test fragments at a Tanimoto similarity threshold of 0.7. Attributions are reported as mean ± std across five seeds.

Dataset	Group	Median atoms	Mean MUG ($\times 10^{-4}$)	Mean MUG / atom ($\times 10^{-5}$)
BBBP	Exact seen	3	0.697 ± 0.112	1.789 ± 0.324
	Similar seen	11	1.701 ± 0.149	1.330 ± 0.153
	Unseen	10	4.884 ± 0.977	8.901 ± 1.794
CYP3A4	Exact seen	3	0.986 ± 0.242	2.475 ± 0.582
	Similar seen	13	4.456 ± 1.430	5.896 ± 2.581
	Unseen	13	3.887 ± 0.925	3.392 ± 0.812
Caco2	Exact seen	3	0.104 ± 0.004	0.229 ± 0.007
	Similar seen	9	0.316 ± 0.023	0.336 ± 0.035
	Unseen	16.5	0.464 ± 0.023	0.295 ± 0.014
Lipophilicity	Exact seen	4	0.192 ± 0.018	0.490 ± 0.057
	Similar seen	10	0.516 ± 0.044	0.555 ± 0.044
	Unseen	11	0.745 ± 0.064	0.745 ± 0.070
AqSolDB	Exact seen	3	0.169 ± 0.015	0.431 ± 0.026
	Similar seen	10	0.777 ± 0.129	0.609 ± 0.108
	Unseen	12	0.987 ± 0.141	0.721 ± 0.097
Tox21	Exact seen	3	0.474 ± 0.071	1.074 ± 0.158
	Similar seen	8	1.620 ± 0.249	2.023 ± 0.276
	Unseen	11	2.904 ± 0.550	2.667 ± 0.568

F.6 Details of MUG Highlights Training-Unseen Substructures

Dataset-level results are provided in Table 25. Table 26 presents a more detailed analysis that introduces an additional category, *similar seen*, comprising fragments that do not occur exactly in the training set but have a Tanimoto similarity of at least 0.7 to a training fragment. For the non-normalized scores, we observe the ordering *exact seen* $<$ *similar seen* $<$ *unseen* in five of the six datasets. This result supports the hypothesis that the model assigns higher uncertainty contributions to less familiar structures. For the atom-normalized scores, however, this ordering does not hold in three datasets. In BBB Martins, similar-seen fragments receive lower attributions than exact-seen fragments, whereas in CYP3A4 Veith and Caco2 Wang, unseen fragments receive lower attributions than similar-seen fragments.

F.7 Details Of Scope And Limits Of The Completeness Guarantee

Table 27 empirically delineates the scope of Theorem 1. Homogeneous readouts and heads yield small residuals of 10^{-6}–10^{-5}, consistent with finite-precision numerical error, whereas attention pooling, biased ReLU heads, input-level attribution, independently biased ensemble members, and probability-space targets violate the theorem's assumptions and produce substantially larger residuals. By contrast, changing the ensemble size or variance normalization does not affect completeness. These results also emphasize that completeness and occlusion fidelity are distinct properties: a method may rank influential atoms effectively even when its attributions do not form an exact decomposition of predictive variance.

Table 27: **Ablations with a GCN backbone.** Results are averaged over the six tasks, except for the probability target, which is evaluated over the three classification tasks. Rows marked † are those where Theorem 1 *predicts* a nonzero residual; every other row should sit at numerical precision.

Ablation	Setting	Residual ↓	ΔFid ↑
Readout ρ (A1)	mean pool	1.26×10^{-5}	0.158
	sum pool	4.62×10^{-6}	0.122
	max pool	1.08×10^{-5}	0.098
	attention pool†	2.96×10^{-1}	0.022
Head ψ (A2)	linear	1.27×10^{-5}	0.158
	bias-free ReLU MLP	1.24×10^{-5}	0.286
	ReLU MLP with biases†	6.60×10^{-2}	0.305
Attribution layer	last GCN layer	1.26×10^{-5}	0.158
	input × gradient†	1.96×10^{-1}	0.316
Ensemble size T	$T = 5$	9.90×10^{-6}	0.073
	$T = 10$	8.76×10^{-6}	0.112
	$T = 30$	1.25×10^{-5}	0.158
	$T = 100$	8.10×10^{-6}	0.183
Uncertainty source (A5)	MC dropout (shared head)	1.25×10^{-5}	0.158
	deep ensemble, per-member bias†	4.85×10^{-2}	0.386
	deep ensemble, tied bias	5.39×10^{-5}	0.402
	deep ensemble, bias-free head	4.46×10^{-5}	0.402
Variance normalization	$1/(T-1)$	1.30×10^{-5}	0.158
	$1/T$	1.30×10^{-5}	0.158
Attribution target (classification)	logit	1.25×10^{-5}	0.158
	probability†	1.04	402.781

Table 28: **Negative attributions and completeness residuals versus the number of MC-dropout passes.** Results are six-task macro-averages using a GCN backbone and scaffold splits. Within each dataset, we compute the fraction of test atoms with strictly negative MUG attributions and the fraction of molecules containing at least one such atom. For each method, we average the molecule-level completeness residual $r(\mathcal{G})$ (Eq. 18) within each dataset, then average equally across datasets. Under the mean-pooling and linear-head architecture used here, Grad-CAM applied to predictive variance equals $\mathrm{ReLU}(\mathrm{MUG})$.

T	Negative atoms (%)	Molecules with ≥ 1 negative atom (%)	Mean MUG residual ↓	Mean Grad-CAM residual ↓
5	27.8 ± 0.2	95.5 ± 0.2	$(2.66 \pm 0.30) \times 10^{-5}$	$(2.93 \pm 0.09) \times 10^{-1}$
10	19.7 ± 0.2	89.9 ± 0.6	$(1.53 \pm 0.10) \times 10^{-5}$	$(9.81 \pm 0.14) \times 10^{-2}$
30	8.9 ± 0.2	69.1 ± 1.0	$(1.90 \pm 0.08) \times 10^{-5}$	$(1.63 \pm 0.03) \times 10^{-2}$
100	2.5 ± 0.1	30.6 ± 1.2	$(1.25 \pm 0.06) \times 10^{-5}$	$(1.56 \pm 0.11) \times 10^{-3}$
200	1.0 ± 0.1	14.5 ± 0.7	$(1.24 \pm 0.07) \times 10^{-5}$	$(3.55 \pm 0.29) \times 10^{-4}$
500	0.3 ± 0.0	4.4 ± 0.6	$(1.63 \pm 0.06) \times 10^{-5}$	$(5.85 \pm 0.84) \times 10^{-5}$
1000	0.1 ± 0.0	1.8 ± 0.3	$(1.54 \pm 0.03) \times 10^{-5}$	$(2.62 \pm 0.22) \times 10^{-5}$
10000	0.0 ± 0.0	0.3 ± 0.2	$(1.18 \pm 0.07) \times 10^{-5}$	$(1.41 \pm 0.05) \times 10^{-5}$

F.8 Why the Rectified Variant Is Not Complete

Applying a ReLU to E_m, as Grad-CAM does, is not innocent here. Lemma 1 gives $E_m^{\sigma^2} \propto \mathrm{Cov}_t(f_t, p_m^{(t)})$ with $p_m^{(t)}$ fragment m's contribution to f_t, so a fragment anti-correlating with the consensus receives a negative share, and clipping makes $\sum_m \mathrm{ReLU}(E_m^{\sigma^2}) \neq \sum_m E_m^{\sigma^2}$ whenever such a share is present. How often this bites depends on T: few passes make the covariances noisy and negative shares common; large T reduces their prevalence in our experiments. We observe negative shares in 95.50% of molecules at $T{=}5$ against 0.3% at $T{=}10000$ (Table 28)—rectification degrades gracefully, but does not preserve the completeness guarantee in general. Vanilla saliency is further removed still: it norms the rectified gradient with respect to the *input* features, discarding sign, collapsing channels, never multiplying by activations.

F.9 Details of application: uncertainty-guided fragment library design

F.9.1 Dataset construction and evaluation protocol

Product library. The study uses a fixed library of acid–amine products derived from Enamine's Essential Fragment Library (Enamine Ltd., 2024). Each library record contains the product SMILES and the SMILES of its two source building blocks: an acid-side fragment A and an amine-side fragment B. The supplied product table contains 13,284 records. We canonicalize the product and building-block SMILES with RDKit and deduplicate by canonical product SMILES, retaining the first record for each product. This yields 13,176 distinct products associated with 91 distinct acid-side and 120 distinct amine-side building blocks. There are 195 distinct building-block structures in total, because 16 occur in both roles. Canonicalization preserves stereochemical information; no additional salt stripping or tautomer standardization is applied.

Docking labels. We dock the library against `5L2S` and `8PJK` using AutoDock Vina 1.2.5 (Eberhardt et al., 2021). Successfully computed docking scores serve as labels in the active-learning simulation and are revealed as compounds are acquired.

Surrogate model and training. All strategies use the same three-layer GCN, with 128 hidden channels per layer. Each graph-convolution layer is followed by dropout with probability 0.2 and a ReLU activation. Global mean pooling and a linear output head produce a scalar docking-score prediction.

At round 0 and after every acquired batch, a new model is trained from scratch on all labels in L_r, with a fresh optimizer and scheduler. Training uses mean squared error on docking scores in their original units, AdamW with learning rate 10^{-3} and weight decay 5×10^{-4}, batch size 32, and 50 epochs. The learning rate is halved after 30 epochs.

Acquisition setup. At round r, let L_r denote the labelled training set and U_r the remaining unlabelled acquisition pool. We sample 500 compounds uniformly without replacement from U_r to form the candidate set C_r, then acquire a batch of 25 compounds. For a compound G, $\mu_r(G)$ and $\sigma_r(G)$ denote the predictive mean and sample standard deviation estimated from $T = 800$ MC-dropout passes.

F.9.2 Acquisition methods

Lower docking scores are preferred. All methods acquire 25 distinct compounds per round.

Random molecules. Select 25 compounds uniformly without replacement from C_r.

Greedy. Select the 25 candidates with the lowest predictive means $\mu_r(G)$.

UCB. Select the 25 candidates with the lowest $\mu_r(G) - \sigma_r(G)$, combining favourable predictions with higher uncertainty.

Shared fragment-based acquisition. Fragment methods visit seed compounds from C_r in a method-specific order and choose an acid-side (A) or amine-side (B) building block. They retrieve compounds with the same canonical building-block SMILES on that side from the full pool U_r. At most 200 available partners are sampled uniformly for ranking, and up to five are acquired per selected fragment. Repeated fragment keys and already selected compounds are skipped. Selection continues until the batch is full. MUG-based methods rank partners by increasing $\mu_r(G)$, whereas Random fragments ranks them randomly.

Random fragments. Visit candidates in random order, choose side A or B with equal probability, and select partners randomly using the shared fragment procedure.

Table 29: **Uncertainty-guided library design for 8PJK.** 50 initial compounds, followed by 50 rounds of 25 acquisitions. Hits score ≤ -9.913 kcal/mol (top 10% of the docked library). Results are mean $\pm$ SD over five seeds. Best means are bold and second-best means are underlined.

Acquisition	Avg. score (kcal/mol) ↓	Best score (kcal/mol) ↓	Hit rate ↑
Random molecules	-8.849 ± 0.008	-11.838 ± 0.137	0.103 ± 0.007
Greedy	$\underline{-9.758 \pm 0.045}$	$\mathbf{-12.360 \pm 0.000}$	$\underline{0.428 \pm 0.020}$
UCB	-9.728 ± 0.028	$\underline{-12.320 \pm 0.055}$	0.416 ± 0.013
Random fragments	-8.807 ± 0.018	-11.900 ± 0.118	0.093 ± 0.010
MUG (random fragment)	-9.718 ± 0.041	$\mathbf{-12.360 \pm 0.000}$	0.422 ± 0.012
MUG	$\mathbf{-9.935 \pm 0.024}$	$\mathbf{-12.360 \pm 0.000}$	$\mathbf{0.523 \pm 0.010}$

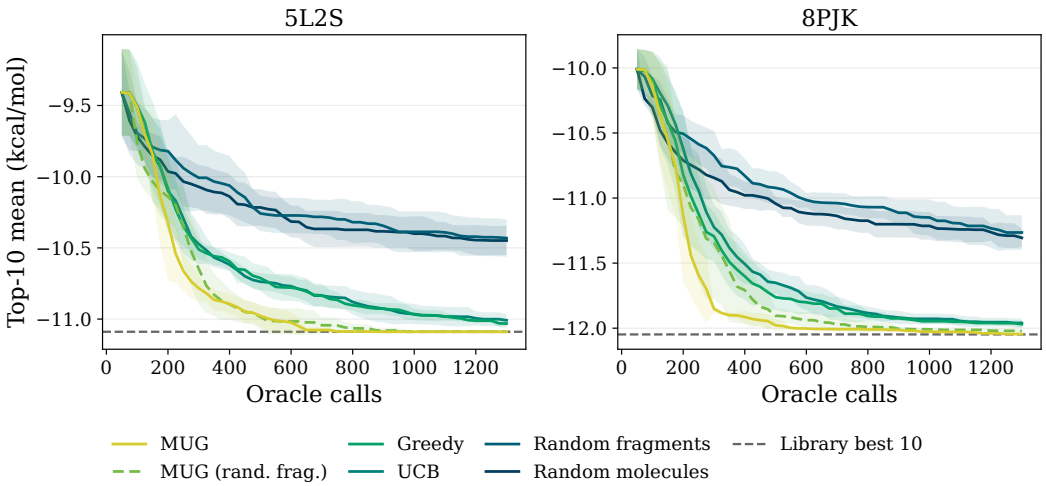


Figure 10: **Mean docking score of the 10 best compounds acquired so far (lower is better).** Mean $\pm$ SD over five seeds. The gray dashed line marks the library's top-10 mean. MUG (rand. frag.) is MUG with the fragment chosen at random.

MUG. Let $E_{r,m}(G)$ be the signed MUG attribution of atom m, $V(G)$ the product's atom set, and $I_F(G)$ the atoms mapped to building block $F \in \{A, B\}$. Define the normalized fragment attribution

$$q_{r,F}(G) = \frac{\sum_{m \in I_F(G)} E_{r,m}(G)}{\sum_{m \in V(G)} E_{r,m}(G)}, \qquad c_r(G) = \max_{F \in \{A,B\}} q_{r,F}(G).$$

Empty mappings are omitted from the maximum; if neither mapping is available, $c_r(G) = 0$. A nonpositive denominator yields zero scores for available fragments.

Let R_r denote ascending ranks among candidates, scaled to $[0, 1]$. MUG visits seeds in descending order of

$$S_r(G) = \tfrac{1}{2} R_r(-\mu_r(G)) + \tfrac{1}{2} R_r(c_r(G)) \,.$$

For each seed, it chooses the available fragment with the larger attribution and acquires partners using the shared procedure.

MUG (random fragment). Use the same seed-ranking and partner-selection rules as MUG, but choose side A or B with equal probability. This ablation isolates the contribution of attribution-guided fragment choice while retaining MUG-based seed ranking.

Additional results. Table 29 reports results for `8PJK`. MUG achieves the best mean docking score and hit rate, increasing the hit rate from 42.2% to 52.3% compared with random fragment choice.

Acquisition trajectories. Figures 10–12 show how the endpoints reported in Tables 6 and 29 develop over the campaign, for both targets and all six acquisition rules.

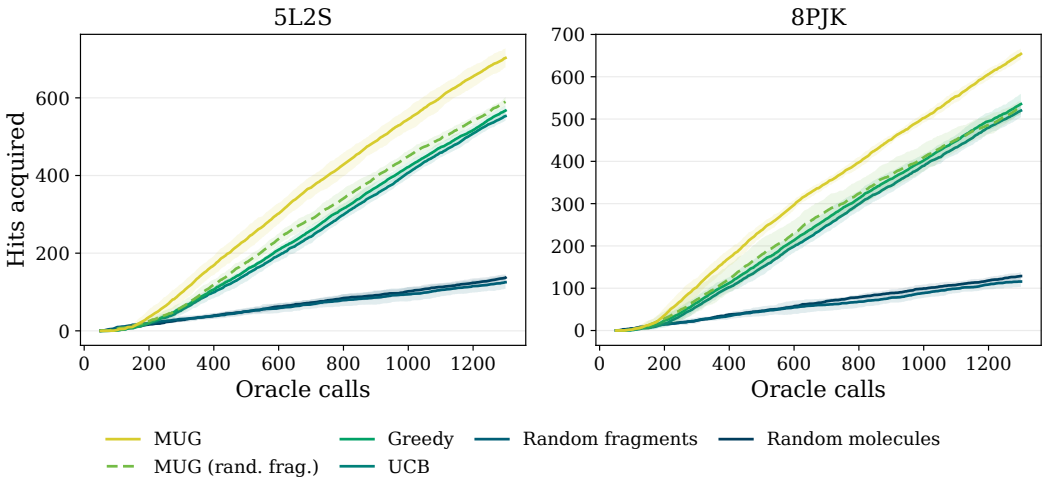


Figure 11: **Cumulative number of hits acquired against oracle calls.** Hits score at or below the top 10% of the docked library (-9.193 kcal/mol for `5L2S`, -9.913 kcal/mol for `8PJK`). Mean $\pm$ SD over five seeds. Dividing the endpoint by the 1,250 acquired compounds recovers the hit rate in Tables 6 and 29.

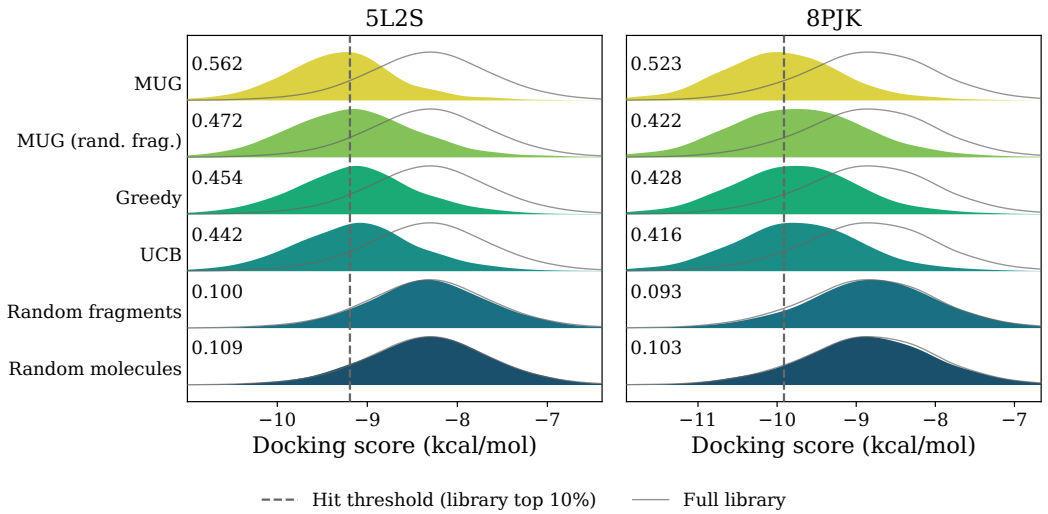


Figure 12: **Docking-score distribution of the 1,250 acquired compounds.** Kernel density estimates pooled over five seeds; the gray outline repeats the full docked library and the dashed line marks the hit threshold. The number on each row is that rule's hit rate.

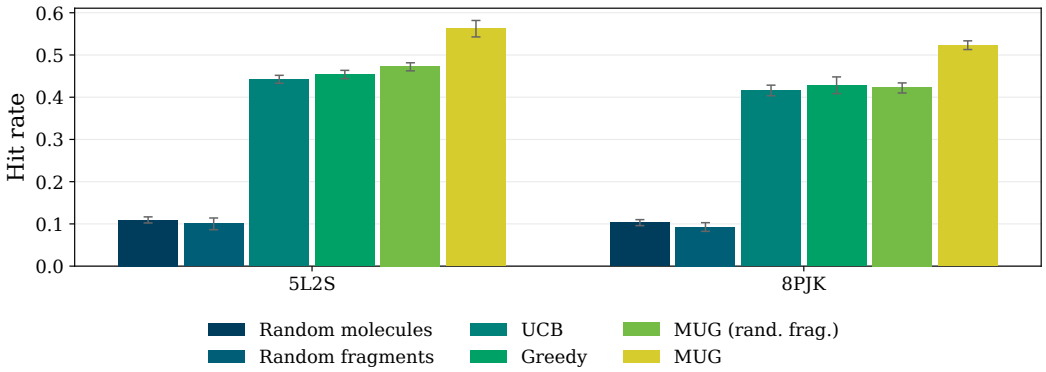


Figure 13: **Hit rate by target.** The hit-rate column of Tables 6 and 29 shown together; mean ± SD over five seeds.

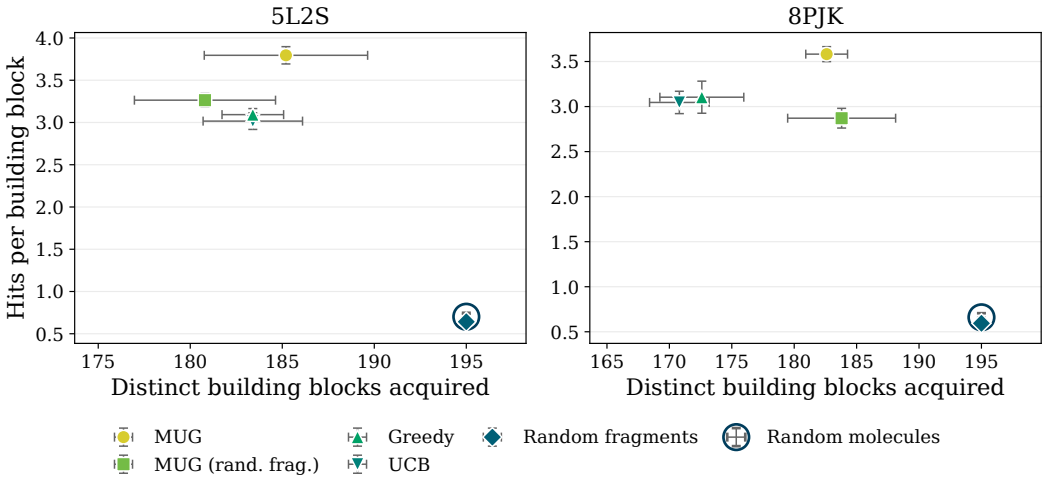


Figure 14: **Fragment economy of the acquired compounds.** Distinct acid- and amine-side building blocks present among the 1,250 acquired compounds, against the number of hits obtained per building block touched. This is defined for every rule, since each product is one acid plus one amine; the non-fragment rules do not steer this count. Mean ± SD over five seeds. The two random rules coincide at all 195 building blocks, so their markers are superimposed (the hollow circle marks Random molecules).