Look Before You Leap: Improving Structure-Based Drug Optimization with Attribution-Guided Genetic Operators

Anonymous Authors[1]

Abstract

Graph-based molecular property predictors are increasingly central to drug discovery, yet the atom-level information they encode remains largely unexplored for guiding molecular optimization. We propose attribution-guided site selection, a modular modification to the Graph-Based Genetic Algorithm (GB-GA) that biases crossover and mutation toward atoms estimated to have the greatest potential to improve predicted fitness, using attribution scores derived from a directed message-passing neural network. On the DOCKSTRING molecular design benchmarks, the three attribution-guided variants tested systematically outperform unmodified GB-GA across targets. The modification is computationally lightweight, requires no retraining, and is orthogonal to outer-loop optimization strategies.

1. Introduction

The discovery and design of molecules with desired biological and chemical properties is a problem of broad scientific and commercial interest, with applications spanning drug discovery, materials design, and reaction catalysis. The vastness of chemical space and the high cost of experimental synthesis and testing make exhaustive exploration infeasible, motivating the development of computational methods for efficient chemical space traversal.

Machine learning models for molecular property prediction are increasingly accurate, but their role in molecular design pipelines has remained largely passive: they score candidate molecules, but the structural information they encode about why a molecule scores well or poorly is discarded after prediction. Attribution methods, which decompose predictions into subgraph-level contributions, have been developed extensively as interpretability tools; however, their potential as operational components that actively guide molecular generation remains underexplored. This gap is relevant as the field moves toward AI systems that not only predict molecular properties but reason about where and how to intervene in molecular structure, a capability central to generative molecular design and emerging agentic approaches to scientific discovery.

We address this gap in the context of the Graph-Based Genetic Algorithm (GB-GA) (Jensen, 2019), an evolutionary method operating directly on 2D molecular graphs that remains competitive with substantially more complex generative methods (Tripp & Hernández-Lobato, 2023; Du et al.,

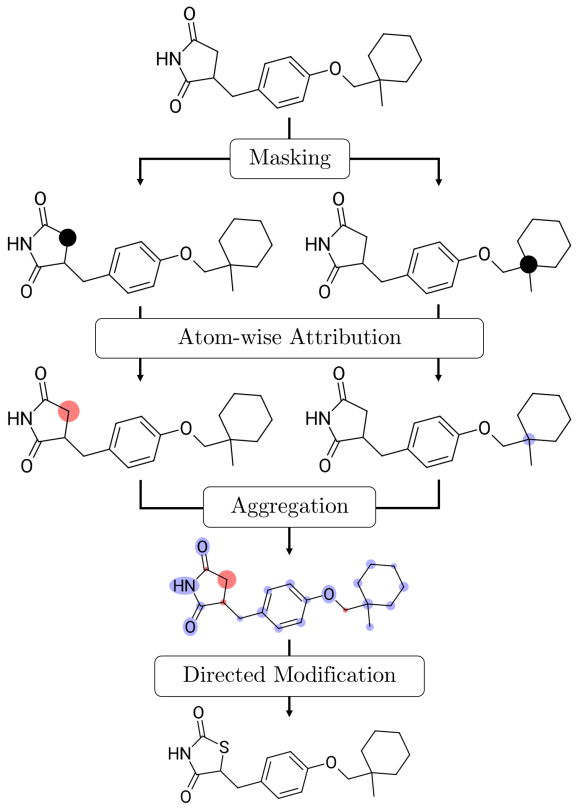

Figure 1. Overview of proposed atom attribution framework.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to/Accepted at/Published in the AI for Science workshop (ICML 2026).

2024). GB-GA's crossover and mutation operations select modification sites uniformly at random, ignoring available information about which atoms contribute to or detract from molecular fitness. We propose attribution-guided site selection, biasing genetic operations toward atoms that a learned property predictor identifies as contributing least to predicted fitness. Our primary contributions are:

- We propose and evaluate three atom-level attribution methods — single-atom ablation, Shapley value attribution, and subgraph removal — as a site-selection mechanism for GB-GA. The modification is modular, preserves standard genetic operators, and is agnostic to the choice of attribution method.
- We demonstrate that attribution-guided variants systematically outperform unmodified GB-GA across all three DOCKSTRING benchmark targets under both favorable and random initialization, with improvement magnitude tracking surrogate model reliability.
- We show that offline surrogate-scored optimization with post-hoc oracle evaluation matches or outperforms direct oracle optimization under the same evaluation budget, approaching Bayesian optimization baselines that require iterative surrogate retraining.

2. Related Work

2.1. Molecular optimization

Computational approaches to molecular design span generative models that learn and sample from distributions over chemical space, including VAEs, diffusion models, and normalizing flows (Du et al., 2024), as well as search-based methods that directly traverse chemical space without learning a generative distribution, including genetic algorithms, reinforcement learning, and Bayesian optimization (Bilodeau et al., 2022). Despite the rapid development of deep generative methods, the Graph-Based Genetic Algorithm (GB-GA) (Jensen, 2019), which performs crossover and mutation directly on 2D molecular graphs, remains competitive with substantially more complex approaches on standard benchmarks (Tripp & Hernández-Lobato, 2023; Du et al., 2024). Relative to generative models, GB-GA requires no pre-training, has low computational overhead, and imposes no distributional assumptions on the molecules it can explore, making it well-suited to low-data regimes and objectives that are difficult to model globally.

2.2. Attribution in molecular models

Attribution methods decompose a model's prediction into per-feature importance scores and have been widely applied to molecular property predictors as post-hoc interpretability tools. General-purpose methods including SHAP (Lundberg & Lee, 2017) and integrated gradients (Sundararajan et al., 2017) have been adapted to the molecular setting, while graph-specific approaches such as GNNExplainer (Ying et al., 2019) and attention-based attribution (Xiong et al., 2020) exploit the node-level structure of GNNs to produce atom- or bond-level importance maps. While these methods have been used almost exclusively for explanation, we repurpose attribution operationally: rather than explaining a fixed prediction, we use atom-level scores to guide where structural modifications are applied to produce a new molecule. This reframing from interpretability to optimization guidance requires no new machinery but changes the role attribution plays in the molecular design pipeline.

While Lee et al. have applied Shapley values to guide a genetic algorithm for protein sequence optimization (2023), extending this principle to small-molecule graphs introduces non-trivial differences: molecular graphs lack the sequential ordering that makes feature attribution straightforward in string representations, requiring perturbations to be defined over unordered node neighborhoods rather than positional sequence elements. This affects both how attribution scores are computed and how candidate operation sites are identified and applied during genetic operations. Though atom-level attribution has been used to define fragment boundaries via aggregation across molecular substructures in molecular optimization (Wüthrich et al., 2021), its use as a modular, probabilistic site-selection bias over standard genetic operators has not been explored. We investigate this formulation systematically, comparing three attribution methods and ablating the role of attribution directionality. To our knowledge, this is the first use of atom attribution to directly guide atom-level site selection in molecular optimization.

2.3. From prediction to intervention

The shift from using learned models as passive scorers to using them as active guides for structural modification connects to a broader trend in AI for science: the development of systems that not only predict properties but reason about where and how to intervene in the objects they model. In molecular generation, this pattern appears in reinforcement learning over molecular edits (Bilodeau et al., 2022), masked pretraining objectives that learn to reconstruct perturbed molecular graphs (Rong et al., 2020), and fragment-based generative methods that condition generation on local structural context. Attribution-guided site selection instantiates this pattern in a minimal form by using a learned model's sensitivity map to direct a search algorithm's modifications, and provides empirical evidence that even simple forms of model-guided intervention improve optimization outcomes. The modularity of this approach suggests that it could serve as a component within more complex AI-driven discovery systems.

3. Methods and Algorithms

3.1. Graph-Based Genetic Algorithm

GB-GA (Jensen, 2019) is an evolutionary algorithm operating directly on 2D molecular graphs. Each generation, parent molecules are paired for crossover, in which each parent contributes a molecular subgraph to produce a child molecule. Offspring may undergo mutation with fixed probability, applying a localized structural modification. The resulting candidates are scored and the top-ranking molecules are retained as parents for the next generation. Crossover and mutation sites are selected uniformly at random in the standard implementation. All generated molecules are guaranteed to be chemically valid.

3.2. Surrogate scoring

We employ a D-MPNN ensemble (Chemprop v2.0 (Heid et al., 2024; Graff et al., 2026)) as the GB-GA fitness function in place of the true oracle. At the conclusion of the run, the highest-scoring unique compounds identified by the surrogate across all generations are submitted for oracle evaluation. A separate single D-MPNN provides atom-level attribution scores; using distinct models for scoring and attribution reduces the risk that attribution scores reflect artifacts of a single model. Training details are provided in the appendix.

3.3. Attribution-guided site selection

Each attribution method produces a score for every atom in a molecule by measuring the effect of removing that atom's contribution to the D-MPNN predicted fitness. For a molecule with n atoms, the $\lfloor n/10 \rfloor$ atoms with the most positive attribution scores (those estimated to contribute least to fitness) are designated as candidate operation sites. Site selection proceeds from this candidate set with bias probability $b = 0.5$, and from the full atom set otherwise, preserving exploratory capacity when attribution scores are uninformative. Sensitivity to b is evaluated in the discussion (Section 5.5). The atom threshold was chosen as a simple heuristic that restricts operations to a small but non-trivial fraction of the molecule, the value of which was not optimized in this work. The consistent improvement across therapeutic targets and attribution methods suggests that the benefit derives from the directionality of attribution guidance rather than the specific threshold, though adaptive scheduling of b or replacing the hard threshold with continuous probability weighting proportional to attribution magnitude are natural refinements.

The three methods differ in how the perturbation is defined. Let f denote the attribution model, G the original molecular graph, V the set of atoms in G, and A_i the attribution score of atom i:

Single-atom ablation sets atom i's feature vector to zero and measures the change in predicted score:

$$A_i = f(\tilde{G}_i) - f(G) \tag{1}$$

where $\tilde{G}_i$ is the graph with atom i's features zeroed. This requires n forward passes per molecule and captures first-order contributions only. The zero-feature baseline does not correspond to a chemical entity but provides a consistent reference when a natural baseline is unavailable (Sundararajan et al., 2017).

Shapley value attribution estimates each atom's average marginal contribution across all possible subsets of co-present atoms (Lundberg & Lee, 2017; Li et al., 2024):

$$A_i = \sum_{S \subseteq V \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[f(G_{S \cup \{i\}}) - f(G_S) \right] \tag{2}$$

where G_S denotes the graph with all atoms outside S zeroed. Because exact computation is intractable, we approximate in this work using $m = 5$ permutation samples, requiring $5n$ forward passes per molecule. Shapley values capture higher-order interactions at greater computational cost.

Subgraph removal structurally removes atom i and its incident edges, evaluating the model on the modified graph:

$$A_i = f(G_{-i}) - f(G) \tag{3}$$

This avoids specifying an arbitrary baseline but can confound attribution with topological disruption for high-degree atoms. Subgraph removal requires n forward passes per molecule.

4. Experiments

4.1. Benchmark

We evaluate on the DOCKSTRING benchmark (García-Ortegón et al., 2022), comprising three de novo molecular design tasks of increasing difficulty: single-target binding (F2), multi-target binding (PPAR), and a selectivity objective with a narrow viable solution space (JAK2). Each objective incorporates protein-ligand binding affinity computed by AutoDock Vina (Trott & Olson, 2010) and a drug-likeness penalty. A reference dataset of 260,000 docked compounds is provided for surrogate training, and all methods are subject to a budget of 5,000 docking oracle evaluations. Objective formulations are provided in the appendix.

4.2. Setup

For each target, a Chemprop ensemble serves as the surrogate fitness function and a separate Chemprop model provides attribution scores, with all methods sharing the same trained models. GB-GA was run for 1,000 generations (population 500, 50 offspring per generation, mutation

Table 1. DOCKSTRING benchmark results. Top-1 and Top-25 oracle scores (mean ± std) across 20 independent runs for proposed methods and across 3 runs for baselines (García-Ortegón et al., 2022) Higher values are better. For proposed variants, **bolded** values indicate statistically significant improvement over the base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. See Section 4.3 for baseline comparability discussion.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Ours (Best Start)	Base GB-GA	10.30±0.26	9.68±0.16	10.29±0.14	9.85±0.07	9.93±0.11	9.52±0.07
	Single-Atom Ablation	10.40±0.24	**9.82±0.11**	10.36±0.11	**9.91±0.05**	9.97±0.15	9.53±0.05
	Shapley Values	**10.53±0.18**	**9.94±0.07**	**10.41±0.18**	**9.90±0.06**	9.94±0.18	9.49±0.06
	Subgraph Removal	10.37±0.24	**9.82±0.13**	**10.37±0.15**	**9.89±0.03**	**10.04±0.19**	**9.56±0.06**
Ours (Random Start)	Base GB-GA	10.07±0.28	9.58±0.18	10.16±0.18	9.75±0.15	9.97±0.33	9.40±0.27
	Single-Atom Ablation	10.23±0.32	**9.70±0.24**	10.24±0.26	9.82±0.20	9.99±0.31	9.45±0.23
	Shapley Values	**10.35±0.44**	**9.82±0.34**	**10.33±0.14**	**9.88±0.10**	10.05±0.23	9.54±0.24
	Subgraph Removal	10.19±0.32	9.69±0.22	**10.32±0.18**	**9.86±0.10**	9.88±0.25	9.43±0.23
Additional baselines	GB-GA	10.33±0.05	9.35±0.06	9.65±0.06	9.06±0.07	9.56±0.01	9.09±0.04
	SELFIES GA	9.51±0.32	8.67±0.04	9.13±0.20	8.58±0.02	9.52±0.11	8.87±0.03
	BO-GP (EI)	10.78±0.38	10.34±0.49	10.89±0.50	10.43±0.33	10.48±0.51	10.03±0.48
	BO-GP (UCB)	10.87±0.17	9.79±0.17	11.13±0.27	10.25±0.06	9.51±0.07	8.96±0.06
	Random (ZINC)	8.62±0.06	7.87±0.04	8.43±0.19	7.79±0.02	8.85±0.28	8.00±0.04

rate 0.01), and the 5,000 highest-scoring unique compounds across all generations were submitted for oracle evaluation. All methods were evaluated across 20 independent runs; results are reported as mean ± standard deviation. Each method was evaluated under two initialization strategies: **best-start**, initialized with the 1,000 highest-scoring reference compounds, and **random-start**, initialized with 1,000 randomly sampled reference compounds.

4.3. Baselines and comparability

We report baseline results from the original DOCKSTRING paper (García-Ortegón et al., 2022) for context, noting that direct comparison is complicated by differences in optimization strategy. In the original work, GB-GA and SELFIES-GA (Nigam et al., 2022) optimize directly against the oracle, consuming evaluation budget with each scored candidate; their random-ZINC initialization is most comparable to our random-start condition. GP-BO (UCB) and GP-BO (EI) use Gaussian process acquisition functions retrained iteratively with oracle feedback — a computationally expensive online adaptation loop unavailable to our offline surrogate — and are initialized with the highest-scoring reference compounds, most comparable to our best-start condition. Random-ZINC samples molecules at random from the ZINC dataset (Irwin et al., 2020), and represents the minimum performance that can be expected for a molecular optimization algorithm. Our primary comparison is between attribution-guided variants and unmodified surrogate-scored GB-GA under matched conditions, isolating the effect of attribution-guided site selection; baseline results provide broader context for the practical competitiveness of the overall framework.

5. Results and Discussion

5.1. Attribution-guided site selection

Attribution-guided variants outperform unmodified GB-GA in point estimates across nearly all comparisons: 17/18 under both best-start and random-start (Table 1). A sign test across all 36 comparisons (34/36 positive; $p < 10^{-6}$) confirms that the improvement is systematic rather than spurious. Under best-start conditions, 11/18 individual comparisons reach statistical significance (one-sided Welch's t-test, $p < 0.05$); under random-start, 7/18 reach significance. Full p-values are reported in Appendix C.4.

Improvement is strongest on F2 and PPAR, where the surrogate models have the highest test set performance, and weakest on JAK2, where the narrow selectivity objective

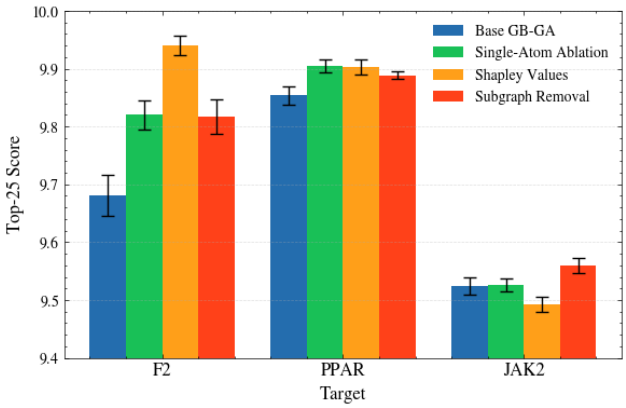

Figure 2. Comparison of top-25 scores (mean ± SEM) under the best-start condition. Higher scores are better.

limits surrogate accuracy outside the training distribution (Appendix C.2). On JAK2, attribution-guided variants consistently achieve higher surrogate-predicted scores than base GB-GA, but this advantage transfers incompletely to oracle evaluation. The failure mode is not in the attribution mechanism itself, as attribution guidance successfully produces higher-scoring candidates under the surrogate, but rather in surrogate-to-oracle transfer: because submitted candidates are selected by surrogate rank, surrogate unreliability at the top of the predicted distribution propagates directly into inconsistent oracle outcomes.

The observed improvements of 0.1–0.3 units in mean Top-25 score are obtained at minimal additional computational cost. Considering the GA run alone, single-atom ablation and subgraph removal require 1.2× the wall-clock time of base GB-GA, while Shapley value attribution requires 2×; however, since docking dominates total pipeline runtime by a factor of 4–17× over the GA step, the end-to-end cost increase is negligible (Appendix C.3). Since the scoring functions combine a docking term with a QED penalty, a unit change does not map cleanly onto binding affinity alone; nonetheless, a 0.2 kcal/mol improvement in binding free energy corresponds to approximately a 40% increase in binding affinity at physiological temperature, a meaningful improvement in a lead-optimization scenario. For further context, in the F2 reference dataset of 260,000 molecules, only 3 other compounds score within 0.2 units of the top-scoring reference molecule, illustrating that absolute gains that appear modest correspond to substantial jumps in rank. That these improvements are achieved without consuming additional oracle evaluations — each corresponding to an expensive docking calculation or experimental assay — makes the cost-benefit tradeoff particularly favorable in practical molecular optimization settings where evaluation budget is the primary constraint.

5.2. Comparison to Other Methods

Attribution-guided variants narrow the gap with GP-BO methods that benefit from iterative surrogate retraining, approaching GP-BO (UCB) on F2 and PPAR and exceeding it on JAK2 (Table 1). GP-BO (EI) achieves the highest top-1 scores on some targets but with substantially higher variance. Notably, each GP-BO iteration involves retraining an acquisition function, requiring up to an hour per round repeated at least 50 times per run based on the reported experimental setup; attribution-guided site selection achieves comparable performance without this overhead. Under random-start conditions, surrogate-scored methods consistently outperform oracle-scored baselines on 5/6 metric-target combinations, with the single exception of F2 Top-1, where oracle-scored GB-GA sometimes achieves a marginally higher point estimate. The consistent advantage on Top-25 metrics suggests that surrogate-guided exploration produces more reliably high-quality candidate sets, even if peak performance on individual targets is occasionally comparable.

5.3. Exploration-exploitation dynamics

Surrogate-predicted score trajectories reveal that the benefit of attribution guidance depends on optimization phase. Under best-start conditions, attribution-guided variants outperform base GB-GA throughout optimization, consistent with a population already concentrated near promising regions where targeted modification is immediately productive (Figure 3). Under random-start conditions, base GB-GA initially achieves higher surrogate scores, reflecting the value of undirected exploration when the population is broadly distributed; however, attribution-guided variants surpass it as the population converges under selection pressure (Figure 4). This crossover is consistent with an exploration-exploitation tradeoff: early diversity is valuable for identifying promising regions, while targeted modification becomes advantageous once those regions have been located.

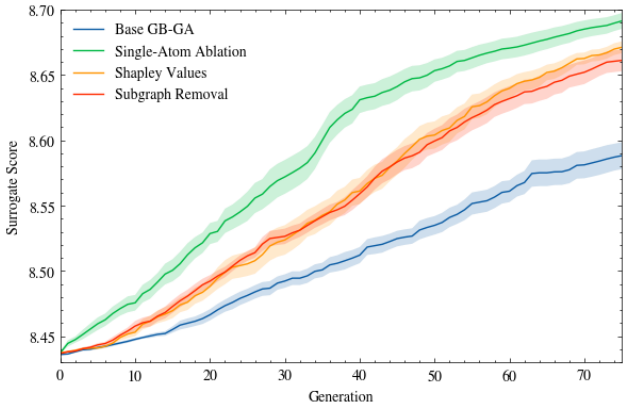

Figure 3. JAK2 surrogate score trajectory (mean ± SEM) under the best-start condition.

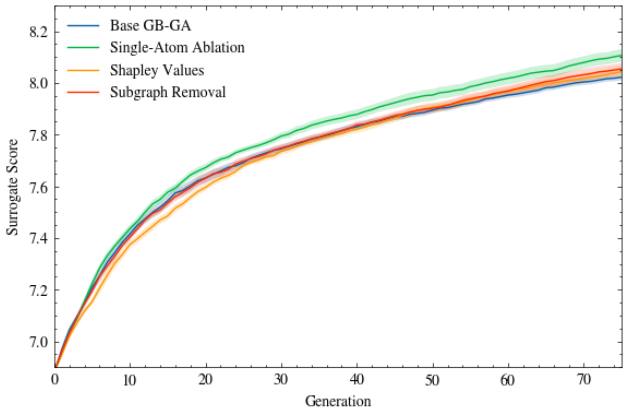

Figure 4. JAK2 surrogate score trajectory (mean ± SEM) under the random-start condition.

5.4. Generated compound diversity

Attribution-guided variants show a target-dependent increase in mean pairwise Tanimoto similarity among top-scoring compounds (Appendix C.5). On F2, similarity is unchanged relative to base GB-GA; on PPAR and JAK2, attribution-guided variants produce moderately more similar populations (mean similarity 0.48–0.55 vs. 0.40), likely reflecting concentration toward narrower high-scoring regions of the surrogate landscape. Absolute similarity values remain well below thresholds associated with trivial analogue generation, but this tradeoff should be considered in applications where scaffold diversity is explicitly valued (Patterson et al., 1996). In practice, the exploitation bias of attribution guidance is well-suited to lead optimization, where the goal is to refine within an established chemical series; for hit identification, where broad chemotype diversity is prioritized, this reduced diversity may require careful attention.

Examining all 5,000 compounds submitted for oracle evaluation per run, attribution-guided variants explore fewer unique molecular scaffolds, consistent with the interpretation that attribution guidance concentrates structural modifications within promising chemical series rather than broadly exploring scaffold space (Appendix C.6). This decrease varies in magnitude across targets, and is most pronounced for Shapley values on F2 (24%) while being modest across PPAR (3-6%). This tradeoff of narrower exploration in exchange for higher-quality candidates within explored scaffolds is consistent with the exploitation-biased role of attribution guidance observed in the trajectory analysis.

5.5. Ablation: Sensitivity to b

Attribution-guided variants were tested on the best-start condition for each of $b \in \{0.25, 0.5, 0.75\}$ under the best-start condition. Performance increases relative to base GB-GA are consistently observed across all bias probabilities tested, with $b \in \{0.5, 0.75\}$ tending to outperform $b = 0.25$ (Appendix D.1).

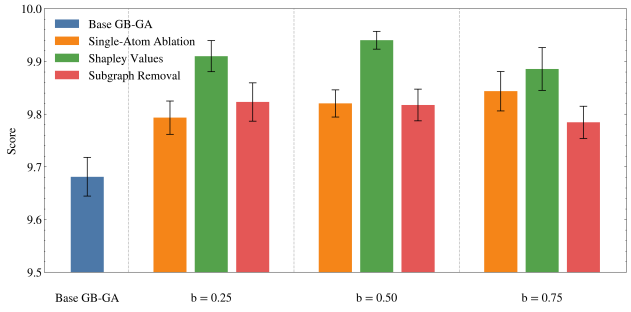

Figure 5. Performance comparison on F2 (mean ± SEM) for different values of b using the best-start condition.

5.6. Ablation: Inverted attribution

To validate that the informational content of attribution scores — not merely non-uniform site selection — drives the observed improvement, we evaluated variants with inverted attribution scores, biasing operations toward atoms estimated to contribute *most* to predicted fitness, under the best-start condition. This serves as an alternative non-uniform site selection strategy that uses the same magnitude information but incorrect directionality. Inverted variants outperform base GB-GA but underperform their corresponding normal-sign variants across targets (Appendix D.2). This ordering — normal attribution > inverted attribution > uniform random — indicates that both directions of informed site selection provide benefit over random selection, but that preferentially modifying low-contributing atoms is the more effective strategy. We interpret this as follows: both normal and inverted attribution concentrate modifications on atoms with high absolute attribution magnitude — atoms that the model identifies as most relevant to predicted fitness — rather than wasting operations on atoms the model considers unimportant. Normal attribution outperforms inverted attribution because modifying atoms that currently detract from fitness has a higher expected payoff than modifying atoms that currently support it.

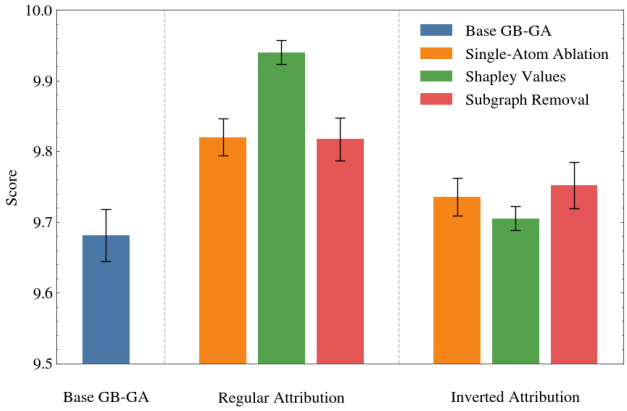

Figure 6. Performance comparison on F2 (mean ± SEM) under normal and inverted attribution under the best-start condition.

6. Conclusion

We propose attribution-guided site selection as a simple, modular modification to GB-GA for molecular optimization, integrating atom-level property attribution with a genetic algorithm framework for the first time. Evaluated on the DOCKSTRING de novo generation benchmarks, the proposed methods achieve competitive performance relative to established baselines, with surrogate-scored optimization consistently matching or outperforming direct oracle optimization under the same evaluation budget. The pro-

posed methods are computationally lightweight, require no modification to the underlying genetic operators, and are orthogonal to outer-loop optimization strategies such as GP-BO, suggesting that attribution-guided site selection is a practical and composable tool for molecular optimization. We further demonstrate that a D-MPNN surrogate trained on reference data can provide meaningful chemical signal for guiding structural operations without access to the true evaluation function, and that surrogate-scored optimization can match or exceed oracle-scored optimization under identical evaluation budgets, which suggests that surrogate smoothing may be broadly beneficial in noisy docking settings.

7. Limitations

This work represents a proof-of-concept investigation of attribution-guided site selection, and several limitations should be noted. The benefit of biased site selection depends on surrogate model reliability, the quality of the starting population, and the structure of the optimization landscape; a principled understanding of when attribution-guided site selection is likely to be beneficial remains an open question. The site selection strategy was not systematically tuned: dynamic adjustment of the bias probability as the population converges, or continuous weighting of site selection by attribution score rather than a hard threshold, are possible avenues for future improvement. Whether the approach remains beneficial when surrogate and attribution models are trained on substantially smaller datasets is an important open question given the practical constraints of many drug discovery settings. The current work uses an offline approach to train models; a natural direction for improvement would be to actively train the surrogate and attribution models, enabling more accurate generalization to out-of-distribution generated molecules. Finally, more expressive architectures such as 2D or 3D graph transformers may produce more reliable atom-level attribution scores; extending the framework to such models is a natural direction for future work, provided they support atom-level attribution.

8. Code Availability Statement

Code and data will be made available upon publication.

9. Impact Statement

This work introduces methods for improving the efficiency of computational molecular optimization, with potential applications in drug discovery and related fields. The proposed methods reduce the number of expensive oracle evaluations required to identify high-scoring molecules, which in the context of drug discovery corresponds to reducing the computational or experimental cost of identifying potent candidate compounds. We do not anticipate direct negative societal impacts from this work. As with all methods that accelerate molecular design, the techniques presented here could in principle be applied to the optimization of harmful compounds; however, the methods themselves are general-purpose optimization tools and do not provide capabilities beyond those already available in the molecular optimization literature. Responsible use of molecular optimization methods in practice requires appropriate domain oversight, which is outside the scope of this work.

References

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. Nature chemistry, 4(2):90–98, 2012.

Bilodeau, C., Jin, W., Jaakkola, T., Barzilay, R., and Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. WIREs Computational Molecular Science, 12(5): e1608, 2022. doi: https://doi.org/10.1002/wcms.1608. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1608.

Bougarne, N., Weyers, B., Desmet, S. J., Deckers, J., Ray, D. W., Staels, B., and De Bosscher, K. Molecular actions of ppar α in lipid metabolism and inflammation. Endocrine reviews, 39(5):760–802, 2018.

Dahlbäck, B. Blood coagulation and its regulation by anticoagulant pathways: genetic pathogenesis of bleeding and thrombotic diseases. Journal of internal medicine, 257(3):209–223, 2005.

Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., and Blundell, T. L. Machine learning-aided generative molecular design. Nature Machine Intelligence, 6(6):589–604, Jun 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00843-5. URL https://doi.org/10.1038/s42256-024-00843-5.

García-Ortegón, M., Simm, G. N. C., Tripp, A. J., Hernández-Lobato, J. M., Bender, A., and Bacallado, S. Dockstring: Easy molecular docking yields better benchmarks for ligand design. Journal of Chemical Information and Modeling, 62(15):3486–3502, 2022. doi: 10.1021/acs.jcim.1c01334. URL https://doi.org/10.1021/acs.jcim.1c01334. PMID: 35849793.

Graff, D. E., Morgan, N. K., Burns, J. W., Doner, A. C., Li, B., Li, S.-C., Manu, J., Menon, A., Pang, H.-W., Wu, H., Zalte, A. S., Zheng, J. W., Coley, C. W., Green, W. H., and Greenman, K. P. Chemprop v2: An efficient, modular machine learning package for chemical property prediction. Journal of Chemical Information and

Modeling, 66(1):28–33, 2026. doi: 10.1021/acs.jcim.5c02332. URL https://doi.org/10.1021/acs.jcim.5c02332. PMID: 41453060.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: A machine learning package for chemical property prediction. Journal of Chemical Information and Modeling, 64(1):9–17, 2024. doi: 10.1021/acs.jcim.3c01250. URL https://doi.org/10.1021/acs.jcim.3c01250. PMID: 38147829.

Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., and Sayle, R. A. Zinc20—a free ultralarge-scale chemical database for ligand discovery. Journal of Chemical Information and Modeling, 60(12):6065–6073, 2020. doi: 10.1021/acs.jcim.0c00675. URL https://doi.org/10.1021/acs.jcim.0c00675. PMID: 33118813.

Jensen, J. H. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. Chem. Sci., 10:3567–3572, 2019. doi: 10.1039/C8SC05372C. URL http://dx.doi.org/10.1039/C8SC05372C.

Lee, B., Seif, Y., Teng, K., Xiao, X., Verma, I., Chen, M.-T., and Cheng, A. C. Ab-deepGA: A generative modeling framework leveraging deep learning for antibody affinity tuning. In NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development, 2023. URL https://openreview.net/forum?id=OrET0hHwLy.

Li, S.-C., Wu, H., Menon, A., Spiekermann, K. A., Li, Y.-P., and Green, W. H. When do quantum mechanical descriptors help graph neural networks to predict chemical properties? Journal of the American Chemical Society, 146(33):23103–23120, 2024. doi: 10.1021/jacs.4c04670. URL https://doi.org/10.1021/jacs.4c04670. PMID: 39106041.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Nigam, A., Pollice, R., and Aspuru-Guzik, A. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. Digital Discovery, 1:390–404, 2022. doi: 10.1039/D2DD00003B. URL http://dx.doi.org/10.1039/D2DD00003B.

Patterson, D. E., Cramer, R. D., Ferguson, A. M., Clark, R. D., and Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. Journal of Medicinal Chemistry, 39(16):3049–3059, 1996. doi: 10.1021/jm960290n. URL https://doi.org/10.1021/jm960290n. PMID: 8759626.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. Self-supervised graph transformer on large-scale molecular data, 2020. URL https://arxiv.org/abs/2007.02835.

Staels, B. Pparγ and atherosclerosis. Current medical research and opinion, 21(sup1):S13–S20, 2005.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

Tripp, A. and Hernández-Lobato, J. M. Genetic algorithms are strong baselines for molecule generation, 2023. URL https://arxiv.org/abs/2310.09267.

Trott, O. and Olson, A. J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, 31 (2):455–461, 2010. doi: https://doi.org/10.1002/jcc.21334. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334.

Vainchenker, W., Leroy, E., Gilles, L., Marty, C., Plo, I., and Constantinescu, S. N. Jak inhibitors for the treatment of myeloproliferative neoplasms and other disorders. F1000Research, 7:82, 2018.

Wüthrich, P., Choong, J. J., and Yuki, S. Infrag: Using attribution-based explainability to guide deep molecular optimization. ChemRxiv, 2021 (0913), 2021. doi: 10.26434/chemrxiv-2021-qtq8d. URL https://chemrxiv.org/doi/abs/10.26434/chemrxiv-2021-qtq8d.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., and Zheng, M. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. Journal of Medicinal Chemistry, 63(16):8749–8760, 2020. doi: 10.1021/acs.jmedchem.9b00959. URL https://doi.org/10.1021/acs.jmedchem.9b00959. PMID: 31408336.

Ying, R., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks, 2019. URL https://arxiv.org/abs/1903.03894.

A. GB-GA and Attribution Details

The implementation of GB-GA in this work is identical to that used in DOCKSTRING (García-Ortegón et al., 2022), with a single modification: site selection for both crossover and mutation is biased toward atoms with high attribution scores, with fixed bias probability b. For crossover, attribution scores bias the choice of fragmentation site; a bond incident to the selected atom is then chosen uniformly at random. For mutation, they bias the atom at which the structural modification is applied.

Specifically, the $\lfloor n/10 \rfloor$ atoms with the highest attribution scores (minimum 1) are selected as candidate operation sites, where n is the number of atoms in the molecule. In practice, this yields 2-4 candidate sites per molecule. With bias probability b, the operation site is drawn from this candidate set; otherwise, it is drawn uniformly from all atoms. Exploration of alternative weighing schemes is left to future work.

B. DOCKSTRING Benchmark Details

Each task in DOCKSTRING involves optimizing an objective function incorporating protein-ligand binding affinity as computed by AutoDock Vina (Trott & Olson, 2010), a widely used molecular docking program. A reference dataset of approximately 260,000 compounds docked against each target is provided for surrogate model training. In the oracle-scored condition, candidate compounds are subject to a budget of 5,000 oracle calls.

Because GB-GA maximizes by default while the original objectives are defined for minimization, we negate all objectives; all reported metrics, including baselines, reflect this convention. Let $s(l, t)$ denote the Vina docking score for ligand l against target t, and $\mathrm{QED}(l)$ the quantitative estimate of drug-likeness (Bickerton et al., 2012). The three de novo design objectives are described below.

B.1. Penalized F2

F2 (prothrombin) is a serine protease central to the coagulation cascade; its inhibition is of therapeutic interest in thrombotic disorders (Dahlbäck, 2005). The F2 objective rewards strong binding to a single target:

$$\mathcal{F}_{\mathrm{F2}}(l) = -\big(s(l, \mathrm{F2}) + 10(1 - \mathrm{QED}(l))\big) \tag{4}$$

B.2. Promiscuous PPAR

The peroxisome proliferator-activated receptors (PPARs) are nuclear receptors involved in regulating lipid metabolism, inflammation, and cellular proliferation (Bougarne et al., 2018). Promiscuous PPAR agonists binding across multiple subtypes have been investigated for treating metabolic syndrome (Staels, 2005). The promiscuous PPAR objective rewards molecules that bind strongly to all PPAR family members:

$$\mathcal{F}_{\mathrm{PPAR}}(l) = -\left(\max_{t \in \{\mathrm{PPARA},\, \mathrm{PPARD},\, \mathrm{PPARG}\}} s(l, t) + 10(1 - \mathrm{QED}(l))\right) \tag{5}$$

B.3. Selective JAK2

Janus kinase 2 (JAK2) is a cytoplasmic tyrosine kinase involved in cytokine signaling, with established roles in myeloproliferative disorders and inflammatory disease (Vainchenker et al., 2018). The selective JAK2 task requires strong binding to JAK2 while avoiding binding to LCK, a lymphocyte-specific kinase whose inhibition is associated with immunosuppressive off-target effects. Kinase selectivity of this form is a prototypical multi-parameter optimization problem, analogous to real-world drug design scenarios in which potency must be balanced against off-target activity, often within a narrow viable solution space. The objective is as follows:

$$\mathcal{F}_{\mathrm{JAK2}}(l) = -\big(s(l, \mathrm{JAK2}) - \min(s(l, \mathrm{LCK}) + 8.1,\ 0) + 10(1 - \mathrm{QED}(l))\big) \tag{6}$$

The penalty term $\min(s(l, \mathrm{LCK}) + 8.1,\ 0)$ is zero when the LCK docking score is weaker (less negative) than the threshold of -8.1, and becomes increasingly negative as LCK binding strengthens beyond this threshold. The objective function as implemented in the released DOCKSTRING code differs from the formula reported in the original manuscript but matches the reported metrics (García-Ortegón et al., 2022); we follow the code implementation as the authoritative version.

C. DOCKSTRING Experiments and Results

C.1. Model training details

The reference dataset, which contains the same compounds for all targets, was clustered by Bemis–Murcko scaffold and split in an 11:1 ratio into training–validation and test sets. The training–validation set was then randomly split four times in a 10:1 ratio, resulting in four distinct training and validation sets.

Hyperparameter optimization was performed for 20 epochs on the F2 objective using the following search space:

- `message-hidden-dim`: {4096, 2048, 1024, 512}
- `ffn-hidden-dim`: {4096, 2048, 1024, 512}
- `depth`: {9, 8, 7, 6}
- `ffn-num-layers`: {6, 5, 4, 3}

All other hyperparameters were left at their default values.

The selected hyperparameters, which were subsequently used for all objectives, were as follows:

- `message-hidden-dim`: 4096
- `ffn-hidden-dim`: 4096
- `depth`: 9
- `ffn-num-layers`: 4

One model was trained on each of the four data splits. Three of these models were used in an ensemble as the surrogate scorer, while the fourth was used for attribution.

C.2. ChemProp model performance

Below, we report the performance of the surrogate ensemble model and attribution model for each target on the held-out test set.

Table 2. Surrogate model test set performance across targets.

Target	MAE	RMSE	R^2	Pearson r
F2	0.282	0.370	0.961	0.980
PPAR	0.325	0.430	0.950	0.975
JAK2	0.336	0.458	0.950	0.975

Table 3. Attribution model test set performance across targets.

Target	MAE	RMSE	R^2	Pearson r
F2	0.295	0.386	0.957	0.978
PPAR	0.341	0.451	0.945	0.972
JAK2	0.353	0.477	0.946	0.973

C.3. Runtime analysis

Runtimes are reported in hours. Model training was conducted on an NVIDIA A10 instance (24 GB VRAM, 8 CPU processes); GB-GA runs were conducted on the same instance using a single CPU process; docking was performed on a G5.4xlarge instance (128 GB RAM, 8 CPU processes). Model training times reflect training all four Chemprop models and represent a one-time cost per target. GB-GA and docking times are averages across runs. Note that due to the number of proteins involved in the scoring function, F2, JAK2, and PPAR require one, two, and three docking runs per molecule, respectively.

Table 4. Runtime analysis for GB-GA and variants.

STEP	F2	PPAR	JAK2
MODEL TRAINING	6.8	6.4	7.2
GB-GA	5.5	5.7	5.7
GB-GA + SINGLE-ATOM ABLATION	6.6	6.8	6.2
GB-GA + SHAPLEY VALUES	10.4	12.9	11.3
GB-GA + SUBGRAPH REMOVAL	6.5	7.2	6.6
DOCKING	24.1	94.8	61.2

C.4. Surrogate and oracle scores

Table 5. Top-1 and Top-25 surrogate scores (mean ± std) across 20 independent runs. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Best-Start	Base GB-GA	9.12±0.01	8.88±0.06	8.76±0.03	8.70±0.02	9.11±0.00	8.78±0.03
	Single-Atom Ablation	**9.15±0.03** (8.7×10^{-5})	**9.04±0.04** (1.5×10^{-12})	**8.79±0.01** (1.5×10^{-6})	**8.72±0.01** (2.3×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.02** (1.5×10^{-9})
	Shapley Values	**9.15±0.03** (8.8×10^{-5})	**9.05±0.02** (3.3×10^{-15})	**8.79±0.01** (3.9×10^{-6})	**8.72±0.01** (3.6×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.03** (3.8×10^{-8})
	Subgraph Removal	**9.14±0.03** (5.7×10^{-3})	**9.01±0.04** (2.3×10^{-10})	**8.79±0.01** (5.1×10^{-6})	**8.72±0.01** (3.1×10^{-7})	9.11±0.00 (5.0×10^{-1})	**8.85±0.03** (1.2×10^{-9})
Random-Start	Base GB-GA	8.85±0.15	8.73±0.14	8.70±0.03	8.64±0.04	8.55±0.07	8.47±0.07
	Single-Atom Ablation	**8.95±0.11** (1.4×10^{-2})	**8.83±0.10** (8.1×10^{-3})	**8.73±0.04** (1.1×10^{-2})	**8.66±0.04** (1.2×10^{-2})	**8.67±0.10** (7.0×10^{-5})	**8.60±0.10** (1.5×10^{-5})
	Shapley Values	**8.96±0.12** (1.0×10^{-2})	**8.86±0.12** (1.8×10^{-3})	**8.73±0.02** (5.3×10^{-4})	**8.68±0.02** (2.4×10^{-5})	**8.67±0.10** (1.1×10^{-4})	**8.60±0.11** (2.9×10^{-5})
	Subgraph Removal	**8.98±0.13** (3.9×10^{-3})	**8.82±0.10** (1.3×10^{-2})	**8.72±0.03** (1.3×10^{-2})	**8.66±0.03** (1.5×10^{-2})	**8.64±0.09** (1.0×10^{-3})	**8.56±0.08** (1.2×10^{-4})

Table 6. Top-1 and Top-25 oracle scores (mean ± std) across 20 independent runs. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Best-Start	Base GB-GA	10.30±0.26	9.68±0.16	10.29±0.14	9.85±0.07	9.93±0.11	9.52±0.07
	Single-Atom Ablation	10.40±0.24 (1.2×10^{-1})	**9.82±0.11** (1.8×10^{-3})	10.36±0.11 (5.4×10^{-2})	**9.91±0.05** (5.9×10^{-3})	9.97±0.15 (1.6×10^{-1})	9.53±0.05 (4.5×10^{-1})
	Shapley Values	**10.53±0.18** (1.6×10^{-3})	**9.94±0.07** (7.5×10^{-8})	**10.41±0.18** (1.3×10^{-2})	**9.90±0.06** (1.3×10^{-2})	9.94±0.18 (3.9×10^{-1})	9.49±0.06 (9.3×10^{-1})
	Subgraph Removal	10.37±0.24 (1.9×10^{-1})	**9.82±0.13** (3.3×10^{-3})	**10.37±0.15** (4.7×10^{-2})	**9.89±0.03** (2.6×10^{-2})	**10.04±0.19** (1.3×10^{-2})	**9.56±0.06** (4.3×10^{-2})
Random-Start	Base GB-GA	10.07±0.28	9.58±0.18	10.16±0.18	9.75±0.15	9.97±0.33	9.40±0.27
	Single-Atom Ablation	10.23±0.32 (5.6×10^{-2})	**9.70±0.24** (4.7×10^{-2})	10.24±0.26 (1.4×10^{-1})	9.82±0.20 (1.3×10^{-1})	9.99±0.31 (4.6×10^{-1})	9.45±0.23 (2.6×10^{-1})
	Shapley Values	**10.35±0.44** (1.2×10^{-2})	**9.82±0.34** (5.5×10^{-3})	**10.33±0.14** (9.6×10^{-4})	**9.88±0.10** (2.5×10^{-3})	10.05±0.23 (2.0×10^{-1})	9.54±0.24 (5.6×10^{-2})
	Subgraph Removal	10.19±0.32 (1.0×10^{-1})	9.69±0.22 (5.9×10^{-2})	**10.32±0.18** (4.4×10^{-3})	**9.86±0.10** (7.9×10^{-3})	9.88±0.25 (8.5×10^{-1})	9.43±0.23 (3.7×10^{-1})

C.5. Tanimoto similarity analysis

To investigate diversity of the generated molecules, mean pairwise Tanimoto similarity was calculated between 25 molecules with the highest oracle scores for each run. We note that this metric may be subject to bias, since a higher-scoring collection of molecules on a given target is likely to be more chemically similar than an arbitrary set of lower-scoring molecules.

Table 7. Top-25 Mean Tanimoto similarity (mean ± std) across 20 independent runs. For proposed variants, **bolded** values indicate statistically significant increases over base GB-GA (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

Method	F2	PPAR	JAK2
Base GB-GA	0.49±0.06	0.40±0.08	0.40±0.06
Single-Atom Ablation	0.48±0.05 (5.5×10^{-1})	**0.53±0.08** (4.0×10^{-6})	**0.50±0.10** (4.2×10^{-4})
Shapley Values	0.46±0.04 (9.4×10^{-1})	**0.55±0.09** (9.6×10^{-7})	**0.51±0.08** (1.8×10^{-5})
Subgraph Removal	0.46±0.04 (9.3×10^{-1})	**0.53±0.08** (3.4×10^{-6})	**0.48±0.07** (2.1×10^{-4})

C.6. Scaffold diversity analysis

Table 8. Number of unique Bemis-Murcko scaffolds in the 5000 submitted compounds (mean ± std) across 20 independent runs. For proposed variants, **bolded** values indicate statistically significant *decreases* relative to base GB-GA (one-sided Welch's t-test, $p < 0.05$); underscored values indicate *increases* relative to base GB-GA. P-values are included in parentheses.

Method	F2	PPAR	JAK2
Base GB-GA	4060±146	4570±254	3084±217
Single-Atom Ablation	4091±225 (6.9×10^{-1})	4537±202 (3.2×10^{-1})	**2739±207** (4.1×10^{-6})
Shapley Values	**3107±238** (1.9×10^{-16})	4454±261 (8.0×10^{-2})	**2691±148** (5.8×10^{-8})
Subgraph Removal	**3784±332** (1.1×10^{-3})	**4285±207** (2.1×10^{-4})	**2693±203** (4.2×10^{-7})

D. Ablation Experiments

D.1. Sensitivity to bias probability

$b \in \{0.25, 0.50, 0.75\}$ were each tested in 20 independent runs, following the same run parameters as the best-start condition. The same surrogate and attribution models were used for all values of b. Note that the results for $b = 0.50$ here correspond directly to the best-start condition in our primary results.

Table 9. Top-1 and Top-25 surrogate scores (mean ± std) across 20 independent runs across $b \in \{0.25, 0.50, 0.75\}$. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Base	Base GB-GA	9.12±0.01	8.88±0.06	8.76±0.03	8.70±0.02	9.11±0.00	8.78±0.03
b=0.25	Single-Atom Ablation	**9.14±0.03** (1.5×10^{-2})	**8.96±0.06** (7.1×10^{-5})	**8.79±0.02** (5.7×10^{-4})	**8.72±0.01** (7.4×10^{-5})	9.11±0.00 (5.0×10^{-1})	**8.82±0.02** (2.4×10^{-7})
	Shapley Values	**9.14±0.02** (1.7×10^{-2})	**9.02±0.04** (7.4×10^{-11})	**8.79±0.01** (7.7×10^{-6})	**8.72±0.01** (2.1×10^{-6})	9.11±0.00 (5.0×10^{-1})	**8.82±0.03** (7.6×10^{-6})
	Subgraph Removal	9.13±0.03 (1.2×10^{-1})	**8.98±0.05** (1.6×10^{-6})	**8.79±0.02** (3.4×10^{-5})	**8.72±0.01** (1.6×10^{-5})	9.11±0.00 (5.0×10^{-1})	**8.84±0.03** (7.4×10^{-7})
b=0.50	Single-Atom Ablation	**9.15±0.03** (8.7×10^{-5})	**9.04±0.04** (1.5×10^{-12})	**8.79±0.01** (1.5×10^{-6})	**8.72±0.01** (2.3×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.02** (1.5×10^{-9})
	Shapley Values	**9.15±0.03** (8.8×10^{-5})	**9.05±0.02** (3.3×10^{-15})	**8.79±0.01** (3.9×10^{-6})	**8.72±0.01** (3.6×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.03** (3.8×10^{-8})
	Subgraph Removal	**9.14±0.03** (5.7×10^{-3})	**9.01±0.04** (2.3×10^{-10})	**8.79±0.01** (5.1×10^{-6})	**8.72±0.01** (3.1×10^{-7})	9.11±0.00 (5.0×10^{-1})	**8.85±0.03** (1.2×10^{-9})
b=0.75	Single-Atom Ablation	**9.17±0.03** (7.8×10^{-8})	**9.04±0.05** (3.1×10^{-11})	**8.79±0.02** (1.9×10^{-4})	**8.73±0.01** (3.1×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.02** (2.5×10^{-8})
	Shapley Values	**9.15±0.03** (3.4×10^{-4})	**9.04±0.05** (2.4×10^{-11})	**8.80±0.01** (1.6×10^{-6})	**8.73±0.01** (1.4×10^{-9})	9.11±0.00 (5.0×10^{-1})	**8.85±0.03** (4.1×10^{-10})
	Subgraph Removal	**9.14±0.03** (4.4×10^{-3})	**9.00±0.05** (3.2×10^{-8})	**8.80±0.01** (1.0×10^{-6})	**8.72±0.01** (9.9×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.85±0.03** (5.0×10^{-9})

Table 10. Top-1 and Top-25 oracle scores (mean ± std) across 20 independent runs across $b \in \{0.25, 0.50, 0.75\}$. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Base	Base GB-GA	10.30±0.26	9.68±0.16	10.29±0.14	9.85±0.07	9.93±0.11	9.52±0.07
b=0.25	Single-Atom Ablation	10.38±0.23 (1.7×10^{-1})	**9.79±0.14** (1.3×10^{-2})	10.32±0.08 (2.3×10^{-1})	9.89±0.06 (6.2×10^{-2})	9.97±0.14 (1.5×10^{-1})	9.56±0.06 (5.2×10^{-2})
	Shapley Values	**10.53±0.20** (2.0×10^{-3})	**9.91±0.13** (1.0×10^{-5})	10.32±0.17 (2.7×10^{-1})	9.88±0.08 (1.2×10^{-1})	9.98±0.14 (1.1×10^{-1})	9.53±0.08 (4.2×10^{-1})
	Subgraph Removal	10.38±0.20 (1.4×10^{-1})	**9.82±0.16** (4.6×10^{-3})	10.37±0.18 (8.7×10^{-2})	9.88±0.05 (6.9×10^{-2})	9.99±0.12 (5.9×10^{-2})	**9.58±0.05** (2.3×10^{-3})
b=0.50	Single-Atom Ablation	10.40±0.24 (1.2×10^{-1})	**9.82±0.11** (1.8×10^{-3})	10.36±0.11 (5.4×10^{-2})	**9.91±0.05** (5.9×10^{-3})	9.97±0.15 (1.6×10^{-1})	9.53±0.05 (4.5×10^{-1})
	Shapley Values	**10.53±0.18** (1.6×10^{-3})	**9.94±0.07** (7.5×10^{-8})	**10.41±0.18** (1.3×10^{-2})	**9.90±0.06** (1.3×10^{-2})	9.94±0.18 (3.9×10^{-1})	9.49±0.06 (9.3×10^{-1})
	Subgraph Removal	10.37±0.24 (1.9×10^{-1})	**9.82±0.13** (3.3×10^{-3})	**10.37±0.15** (4.7×10^{-2})	**9.89±0.03** (2.6×10^{-2})	**10.04±0.19** (1.3×10^{-2})	**9.56±0.06** (4.3×10^{-2})
b=0.75	Single-Atom Ablation	10.42±0.21 (7.1×10^{-2})	**9.84±0.16** (1.8×10^{-3})	10.31±0.09 (3.3×10^{-1})	9.88±0.05 (1.1×10^{-1})	9.94±0.09 (3.9×10^{-1})	9.51±0.04 (7.5×10^{-1})
	Shapley Values	**10.53±0.22** (2.8×10^{-3})	**9.89±0.18** (3.1×10^{-4})	**10.41±0.18** (1.3×10^{-2})	**9.91±0.05** (4.1×10^{-3})	**10.04±0.17** (1.3×10^{-2})	9.51±0.11 (7.1×10^{-1})
	Subgraph Removal	10.38±0.20 (1.6×10^{-1})	**9.78±0.13** (1.8×10^{-2})	10.38±0.19 (5.3×10^{-2})	**9.91±0.05** (4.4×10^{-3})	**10.04±0.16** (5.7×10^{-3})	9.56±0.07 (7.0×10^{-2})

D.2. Inverted attribution

Inverted attribution was tested on all targets using $b = 0.50$ with 20 independent runs, following the same run parameters as the best-start condition. The same surrogate and attribution models were used in inverted and normal attribution. Note that

the normal attribution results here correspond directly to the best-start condition in our primary results.

Table 11. Top-1 and Top-25 surrogate scores (mean ± std) across 20 independent runs for normal and inverted attribution. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Base	Base GB-GA	9.12±0.01	8.88±0.06	8.76±0.03	8.70±0.02	9.11±0.00	8.78±0.03
Normal Attribution	Single-Atom Ablation	**9.15±0.03** (8.7×10^{-5})	**9.04±0.04** (1.5×10^{-12})	**8.79±0.01** (1.5×10^{-6})	**8.72±0.01** (2.3×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.02** (1.5×10^{-9})
	Shapley Values	**9.15±0.03** (8.8×10^{-5})	**9.05±0.02** (3.3×10^{-15})	**8.79±0.01** (3.9×10^{-6})	**8.72±0.01** (3.6×10^{-8})	9.11±0.00 (5.0×10^{-1})	**8.84±0.03** (3.8×10^{-8})
	Subgraph Removal	**9.14±0.03** (5.7×10^{-3})	**9.01±0.04** (2.3×10^{-10})	**8.79±0.01** (5.1×10^{-6})	**8.72±0.01** (3.1×10^{-7})	9.11±0.00 (5.0×10^{-1})	**8.85±0.03** (1.2×10^{-9})
Inverted Attribution	Single-Atom Ablation	9.13±0.03 (1.1×10^{-1})	**8.92±0.07** (2.3×10^{-2})	8.77±0.03 (5.7×10^{-2})	**8.71±0.01** (4.1×10^{-2})	9.11±0.00 (5.0×10^{-1})	**8.79±0.02** (2.9×10^{-2})
	Shapley Values	9.13±0.02 (1.2×10^{-1})	**8.94±0.04** (2.1×10^{-4})	**8.78±0.02** (9.2×10^{-4})	**8.72±0.01** (2.2×10^{-6})	9.11±0.00 (5.0×10^{-1})	**8.88±0.03** (1.9×10^{-14})
	Subgraph Removal	9.13±0.04 (3.5×10^{-1})	**8.96±0.08** (4.9×10^{-4})	8.76±0.02 (3.3×10^{-1})	**8.71±0.01** (4.4×10^{-3})	9.11±0.00 (5.0×10^{-1})	**8.79±0.01** (2.4×10^{-2})

Table 12. Top-1 and Top-25 oracle scores (mean ± std) across 20 independent runs for normal and inverted attribution. For proposed variants, **bolded** values indicate statistically significant improvement over base GB-GA under matched conditions (one-sided Welch's t-test, $p < 0.05$); underscored values indicate decreases relative to base GB-GA. P-values are included in parentheses.

	Objective	F2		PPAR		JAK2	
	Metric	Top-1	Top-25	Top-1	Top-25	Top-1	Top-25
Base	Base GB-GA	10.30±0.26	9.68±0.16	10.29±0.14	9.85±0.07	9.93±0.11	9.52±0.07
Normal Attribution	Single-Atom Ablation	10.40±0.24 (1.2×10^{-1})	**9.82±0.11** (1.8×10^{-3})	10.36±0.11 (5.4×10^{-2})	**9.91±0.05** (5.9×10^{-3})	9.97±0.15 (1.6×10^{-1})	9.53±0.05 (4.5×10^{-1})
	Shapley Values	**10.53±0.18** (1.6×10^{-3})	**9.94±0.07** (7.5×10^{-8})	**10.41±0.18** (1.3×10^{-2})	**9.90±0.06** (1.3×10^{-2})	9.94±0.18 (3.9×10^{-1})	9.49±0.06 (9.3×10^{-1})
	Subgraph Removal	10.37±0.24 (1.9×10^{-1})	**9.82±0.13** (3.3×10^{-3})	**10.37±0.15** (4.7×10^{-2})	**9.89±0.03** (2.6×10^{-2})	**10.04±0.19** (1.3×10^{-2})	**9.56±0.06** (4.3×10^{-2})
Inverted Attribution	Single-Atom Ablation	10.43±0.30 (1×10^{-1})	9.69±0.12 (4×10^{-1})	10.25±0.12 (8×10^{-1})	9.82±0.05 (9×10^{-1})	**10.04±0.16** (2×10^{-2})	**9.60±0.07** (4×10^{-3})
	Shapley Values	10.35±0.22 (3×10^{-1})	9.68±0.08 (5×10^{-1})	10.35±0.14 (2×10^{-1})	9.89±0.05 (1×10^{-1})	**10.03±0.21** (4×10^{-2})	**9.60±0.05** (2×10^{-3})
	Subgraph Removal	10.43±0.24 (1×10^{-1})	9.75±0.07 (1×10^{-1})	10.30±0.15 (5×10^{-1})	9.86±0.05 (4×10^{-1})	9.95±0.12 (3×10^{-1})	**9.57±0.04** (5×10^{-2})