# Bayesian Optimization for Molecules Should Be Pareto-Aware

**Anonymous Authors**
Paper under double-blind review.

## Abstract

Multi-objective Bayesian optimization (MOBO) provides a principled framework for navigating trade-offs in molecular design. However, its empirical advantages over scalarized alternatives remain underexplored. We benchmark a simple Pareto-based MOBO strategy—Expected Hypervolume Improvement (EHVI)—against a simple fixed-weight scalarized baseline using Expected Improvement (EI), under a tightly controlled setup with identical Gaussian Process surrogates and molecular representations. Across three molecular optimization tasks, EHVI consistently outperforms scalarized EI in terms of Pareto front coverage, convergence speed, and chemical diversity. While scalarization encompasses flexible variants—including random or adaptive schemes—our results show that even strong deterministic instantiations can underperform in low-data regimes. These findings offer concrete evidence for the practical advantages of Pareto-aware acquisition in de novo molecular optimization, especially when evaluation budgets are limited and trade-offs are nontrivial.

## 1 Introduction

The discovery of therapeutic molecules is fundamentally a multi-objective optimization problem: a viable candidate must simultaneously satisfy competing criteria such as potency, safety, and pharmacokinetic properties [28, 10, 13]. Scalarization—where multiple objectives are collapsed into a single score using weighted combinations [27, 19]—remains a widely used strategy due to its compatibility with single-objective optimization pipelines. However, scalarization requires *a priori* knowledge of objective weightings, which are often uncertain, context-dependent, or poorly defined in real-world drug design. Furthermore, any fixed weighting yields only a single point on the Pareto front, necessitating repeated and often redundant optimization runs to recover a diverse set of trade-off solutions.

These limitations motivate Pareto-based multi-objective Bayesian optimization (MOBO) methods that preserve vector-valued structure of the problem and directly seek non-dominated solutions across all objectives [14, 40, 1]. Rather than collapsing objectives, MOBO aims to efficiently approximate the Pareto front by guiding evaluations toward regions that improve coverage. This approach has been shown to recover more chemically diverse and balanced solutions using fewer queries—particularly valuable in low-data, expensive-to-evaluate regimes. Beyond molecular design, MOBO has also demonstrated success across domains including materials science [17, 25], protein engineering [30], and robotics [22, 37].

Despite this growing interest, few empirical studies systemically benchmark Pareto-based MOBO against specific scalarization strategies under controlled conditions. Prior work often defaults to scalarization heuristics [13, 21] without evaluating their performance relative to the dedicated Pareto-based acquisitions. In this study, we present a controlled comparison between a fixed-weight scalarized Bayesian optimization strategy (using Expected Improvement) and a Pareto-based MOBO approach (using Expected Hypervolume Imrprovement). Both are implemented using identical

Submitted to AI for Science workshop (NeurIPS 2025)

Gaussian Process surrogates and molecular representations, isolating the acquisition function as the key difference. By evaluating performance across three `GUACAMOL` benchmark tasks, we highlight how even basic Pareto-based strategies can outperform scalarization in data-constrained molecular discovery scenarios—supporting MOBO as a more robust default for early-stage optimization. To ensure reproducibility, we will release the code and data upon acceptance, with links provided in the final version of the paper.

## 2 Background and Related Work

### 2.1 Multi-Objective Optimization and Pareto Optimality

Multi-objective optimization (MOO) concerns optimizing multiple competing objectives simultaneously. Formally, the goal is find $\mathbf{x}^* \in \mathcal{X}$ that maximizes a vector-valued objective $R(\mathbf{x}) = [R_1(\mathbf{x}), \ldots, R_d(\mathbf{x})]$. In general, no single $\mathbf{x}$ maximizes all objectives when they conflict. Instead, the optimal solutions form the *Pareto set*, comprising all $\mathbf{x}$ for which no other $\mathbf{x}'$ improves all objectives. A point $\mathbf{x}_1$ is said to dominate $\mathbf{x}_2$ if $R_i(\mathbf{x}_1) \geq R_i(\mathbf{x}_2)$ for all $i$ and strictly greater for at least one $i$. The *Pareto front* is the image of the Pareto-optimal set in objective space.

Classical techniques to approximate the Pareto front inlcude evolutionary algorithms such as NSGA-II [8] and MOEA/D [39], which maintain a population of candidate solutions. While effective for exploring diverse fronts, they are sample-inefficient and thus prohibitive when objective evaluations are expensive, as is often the case in scientific design problems [9, 24].

### 2.2 Bayesian Optimization with Gaussian Processes

Bayesian optimization is a sample-efficient framework for optimizing expensive black-box functions. It employs a surrogate model—typically a Gaussian Process ($\mathcal{GP}$)—to model the function's uncertainty and selects new evaluations using an acquisition function $\alpha(x)$ that balances exploration and exploitation [3].

A Gaussian Process ($\mathcal{GP}$) is the most widely used surrogate in BO due to its nonparametric flexibility and closed-form posterior. A $\mathcal{GP}$ prior over functions $f\colon \mathcal{X} \to \mathbb{R}$ is specified by a mean function $m(\mathbf{x})$ (often taken as zero) and a covariance (kernel) function $k(\mathbf{x}, \mathbf{x}')$ [33]. Given observations $\mathcal{D}_n = (\mathbf{x}i, y_i)i = 1^n$ with Gaussian noise $y_i = f(\mathbf{x}_i) + \varepsilon_i, ; \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, the $\mathcal{GP}$ posterior at a candidate $\mathbf{x}$ has predictive distribution:

$$f(x)|\mathcal{D}_n \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$$

where:

$$\mu_n(\mathbf{x}) = \mathbf{k}_n(\mathbf{x})^\top (K_n + \sigma^2 I)^{-1}\mathbf{y}, \quad \sigma_n^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{k}_n(\mathbf{x})^\top (K_n + \sigma^2 I)^{-1}\mathbf{k}_n(\mathbf{x}),$$

with $K_n[i, j] = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{k}_n(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}i)]i = 1^n$ [33]. These expressions quantify both the surrogate mean and its epistemic uncertainty, enabling a principled trade-off in $\alpha(\mathbf{x})$.

BO driven by $\mathcal{GP}$ surrogates has achieved remarkable sample efficiency across applications ranging from engineering design to hyperparameter tuning, often requiring orders of magnitude fewer evaluations than grid search or evolutionary algorithms [3]. In molecular optimization, one must also capture chemical similarity in the kernel choice. A popular choice is the *Tanimoto kernel*[32] on binary molecular fingerprints $x, x' \in [0, 1]^d$, defined as:

$$k_{Tanimoto}(x, x') = \frac{x^\top x'}{||x||_2^2 + ||x'||_2^2 - x^\top x'}$$

which measures the ratio of shared substructures to the total fingerprint bits [34]. When used within a $\mathcal{GP}$, the Tanimoto kernel effectively models structure–property relationships in small-molecule spaces and yields strong predictive performance on tasks like binding affinity and solubility prediction [15, 36].

### 2.3 Scalarization and Pareto-based Optimization Strategies

Methods for multi-objective optimization (MOO) broadly fall into scalarization-based and Pareto-based approaches. Scalarization methods reduce a vector-valued objective $f(x) =$

$(f_1(x), \cdots, f_k(x))$ to a scalar-valued surrogate, enabling the use of well-established single-objective acquisition functions such as Expected Improvement (EI) or Upper Confidence Bound (UCB). A classical formulation is the weighted sum $f_{\text{ws}}(x) = \sum_{i=1}^{k} w_i f_i(x)$ where the weights $w_i$ reflect trade-off preferences. While widely used [9, 23], this approach is only guaranteed to recover Pareto-optimal points when the front is convex. Covering non-convex regions typically requires multiple optimization runs with diverse weight configurations, making the approach computationally expensive and potentially redundant.

To overcome these limitations, alternative scalarization techniques such as the Tchebycheff scalarization have been developed, with $f_{Tchebycheff}(x) = \max_{1 \leq i \leq m}\{\lambda_i(f_i(x) = z_i^*)\}$ where $\lambda_i$ are scaling weights and $z_i^*$ is a reference point (typically the ideal vector of component-wise maxima). This scalarization emphasizes the worst-performing objective, yielding stronger guarantees for recovering solutions on convex fronts [2, 26]. Recent work has shown that only a small set of well-chosen Tchebycheff scalarizations can approximate the entire Pareto front with high fidelity and sample efficiency [26].

A particularly impactful advancement comes from the *hypervolume scalarization framework* introduced by Golovin and Zhang [16], which establishes a formal connection between scalarized objectives and the hypervolume indicator—a gold-standard, Pareto-compliant metric [41]. Specifically, they show that for a suitable distribution $\mathcal{D}_\lambda$ over weight vectors and corresponding scalarization functions $s_\lambda$, the hypervolume of a set $Y \subset \mathbb{R}^d$ with respect to a reference point $z \in \mathbb{R}^d$ can be expressed as $HV_z(Y) = c_k \, \mathbb{E}_{\lambda \sim \mathcal{D}_\lambda} \left[\max_{y \in Y} s_\lambda(y - z)\right]$.

This theoretical result provides a principled mechanism for converting hypervolume maximization into a sequence of scalar optimization subproblems. In practice, it enables provable convergence to the Pareto front using standard Bayesian optimization techniques, such as Thompson Sampling or UCB, by simply sampling a new scalarization $s_\lambda$ at each iteration. Golovin and Zhang [16] derive cumulative hypervolume regret bounds of order $\tilde{O}(T^{-1/2})$, establishing the method as both theoretically grounded and sample-efficient for multi-objective black-box optimization.

In contrast to scalarization, Pareto-based acquisition functions preserve the vectorized nature of objectives and directly aim to improve coverage of the Pareto front. Each objective is modelled independently, often using Gaussian Processes and acquisitions are scored based on their *expected contribution* to the Pareto frontier.

The most widely used Pareto-based acquisition is the Expected Hypervolume Improvement (EHVI) [11, 7], which quantifies the increase in dominated volume achieved by adding a new sample. Formally, for the current approximation set $\mathcal{P}_t$ and reference point $z_i$, the EHVI at candidate point $x$ is given by:

$$\text{EHVI}(x) = \mathbb{E}[HV_z(\mathcal{P}_t \cup \{f(x)\}) - HV_z(\mathcal{P}_\sqcup)]$$

Unlike scalarization, which converts MOO into a single-objective landscape, EHVI retains the multi-dimensional structure of the problem, promoting exploration in underrepresented Pareto regions. Refinements, such as Predictive Entropy Search for Multi-objective Optimization (PESMO) [20], take an information-theoretic view, maximizing expected information gain about the Pareto set. Methods such as DGEMO [24] go further by jointly optimizing for hypervolume improvement and diversity in both design and objective space. By modeling the Pareto front as a piecewise manifold and partitioning it into local diversity regions, DGEMO selects a diverse batch of samples to improve coverage efficiency - an especially useful property in low-data regimes.

## 3 Experimental Setup

Our experiments are designed to assess the practical performance of *Pareto-based acquisition strategies* for multi-objective Bayesian optimization (MOBO) in molecular design tasks. Rather than contrasting MOBO with scalarization broadly—which encompasses a diverse array of formulations including random scalarizations [16, 29]—we focus on a controlled comparison between two representative acquisition strategies: Expected Hypervolume Improvement (EHVI), which explicitly targets Pareto front expansion, and a fixed-weight Expected Improvement (EI), a baseline scalarized acquisition function. We aim to answer the following questions:

1. **Optimality:** Does EHVI discover solutions closer to approximate Pareto front than scalarized EI?
2. **Diversity:** Do molecules selected by EHVI exhibit greater structural diversity?
3. **Trade-offs:** How do EHVI and EI differ in balancing optimality and diversity under fixed BO evaluation budgets?

We obtain positive empirical results across the 3 MPOs.

### 3.1 Benchmark Tasks

We evaluate both methods on three multi-property optimization tasks from `GUACAMOL` [4]: Amlodipine MPO, Fexofenadine MPO, and Perindopril MPO. Each task involves optimizing three molecular properties jointly—target similarity, QED, and either logP, SA score, or molecular weight—thus posing realistic trade-offs encountered in drug discovery pipelines.

### 3.2 Optimization Setup

We adopt a multi-objective Bayesian optimization framework where each molecular property $f_j(m)$ is modeled independently using Gaussian Processes ($\mathcal{GP}$s). Each $\mathcal{GP}$ is equipped with the MinMax kernel, a count-aware generalization of the Tanimoto kernel suitable for Morgan fingerprints. The kernel is defined as:

$$k_{\text{MinMax}}(x, x') = \frac{\sum_i \min(x_i, x'_i)}{\sum_i \max(x_i, x'_i)}$$

This kernel measures structural similarity between molecules encoded as extended-connectivity fingerprints (ECFPs) [34] of radius 3, computed via RDKit using count-based features without truncation. Each molecular property is modeled as a $\mathcal{GP}$ prior $f_j \sim \mathcal{GP}(\mu_j, K_j(x_i, x_q))$. Predictions across all objectives yield independent Gaussian posteriors characterized by predictive means and variances $\vec{\mu}(m)$ and $\vec{\sigma}^2(m)$, respectively. The predictive distribution for each $\mathcal{GP}$ is Gaussian, parameterized by a mean and variance derived from exact kernel matrix computations. These $\mathcal{GP}$s are implemented in a JAX-based framework, `kernel_only_GP`, supporting efficient parallelization and differentiable matrix operations. All model hyperparameters, including amplitude $\alpha = 1.0$ and noise variance $s = 10^{-4}$, are held fixed across trials and methods to ensure consistent modeling assumptions.

We compare two acquisition strategies: Expected Hypervolume Improvement (EHVI), a Pareto-based method that promotes non-dominated frontier expansion, and scalarized Expected Improvement (EI) with fixed weights, which collapses the objectives into a single scalar score. EHVI is computed via Monte Carlo sampling using 1000 draws per candidate molecule. At each optimization step, a single molecule is selected from a fixed candidate pool of 10,000 compounds sampled from the `GUACAMOL` training set. The selected molecule is evaluated, added to the training archive, and the $\mathcal{GP}$ models are updated accordingly. Each run proceeds for 200 optimization rounds.

Experiments are repeated with three different random seeds per method. We report the mean and standard deviation of all evaluation metrics. Given the small number of trials, we assess the consistency and magnitude of observed differences using effect size metrics—Cohen's $d$ [6] and Cliff's Delta [5]. All computations are performed on NVIDIA H100-47 GPUs.

### 3.3 Evaluation Metrics and Performance Indicators

To evaluate the performance of multi-objective optimization across both convergence quality and molecular diversity, we adopted three complementary metrics: Hypervolume Indicator (HVI) [12], the R2 indicator [18] and the #Circles metric [38]. The HVI measures the volume of the objective space dominated by the non-dominated solutions relative to a reference point, capturing both convergence to the Pareto front and diversity of trade-offs explored. A higher HVI value indicates broader and more optimal coverage of the objective space. The $R^2$ indicator quantifies the quality of the Pareto front approximation by comparing it to a fixed set of uniformly distributed reference directions, as adapted from the setup in Jain et al. [21]. Specifically, it computes the augmented Tchebycheff scalarization in each direction: for each reference vector $v$, it calculates the worst-case deviation $\max_i v_i \cdot |u_i - s_i|$, where $u$ is the utopian (ideal) point and $s$ is a candidate solution. It then selects the best such solution

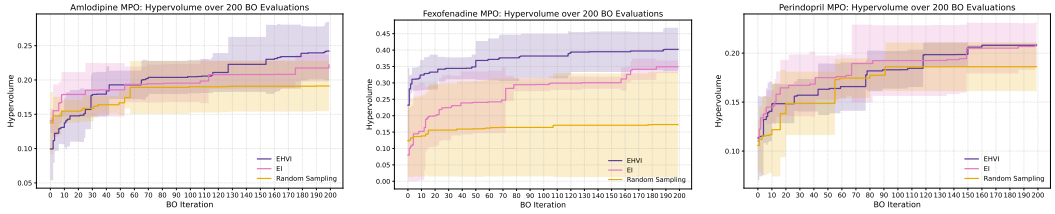
for each direction and averages over all directions. Thus, lower $R^2$ values indicate a front that is both closer and more uniformly distributed relative to the ideal front [18]. Finally, the #Circles metric quantifies structural diversity in the chemical space by counting the number of pairwise dissimilar molecules exceeding a Tanimoto distance threshold $t$. We compute it on the set of Pareto-optimal candidates from the initial and acquired molecules over 200 BO evaluations, as the metric is designed to assess the effective space explored by a *representative* set of high-quality or relevant solutions rather than all generated samples [38]. Informed by axiomatic diversity principles, #Circles offers a geometry-aware, thresholded view of chemical diversity: at higher thresholds (i.e. requiring greater dissimilarity), a larger #Circles value indicates that the candidate set spans more structurally distinct regions of the chemical space.

# 4 Results

This section presents a comparative evaluation of EHVI and scalarized EI across 3 multi-objective molecular optimization tasks. Across all tasks and metrics - hypervolume (Figure 1), $R^2$ indicator (Figure 2), and chemical diversity via the #Circles metric (Figure 3), EHVI demonstrates consistent advantages. Statistical tests provided in Appendix A.2 (Tables 3, 4) corroborate these trends: for hypervolume, EHVI achieves medium to large effect sizes (Cohen's $d = 0.576 - 1.093$) and favorable Cliff's delta values indicating better performance across matched trials. For the $R^2$ indicator, lower values for EHVI suggest improved approximation of the Pareto front, with strong negative effect sizes (e.g. $d = -2.56$ on Fexofenadine MPO). In terms of diversity, EHVI consistently explores more structurally distinct solutions at higher Tanimoto thresholds. These results demonstrate EHVI's robustness: it not only accelerates convergence and enhances front coverage, it also maintains better front approximation and promotes chemically diverse candidate solutions. We next present the detailed metric-wise breakdown across all tasks.

## 4.1 Hypervolume Indicator (HVI)

Across all tasks, EHVI consistently outperforms scalarized EI and random sampling in hypervolume performance (Figure 1. In Amlodipine, EHVI achieves faster convergence and higher final hypervolume with lower variance. For Fexofenadine, the gap is even more pronounced—EHVI dominates throughout, especially in later stages. In Perindopril, both methods reach similar final values, but EHVI converges earlier and exhibits reduced variance, indicating greater sample efficiency and robustness. These results underscore EHVI's consistent advantage in front expansion and reliability across diverse multi-objective settings.



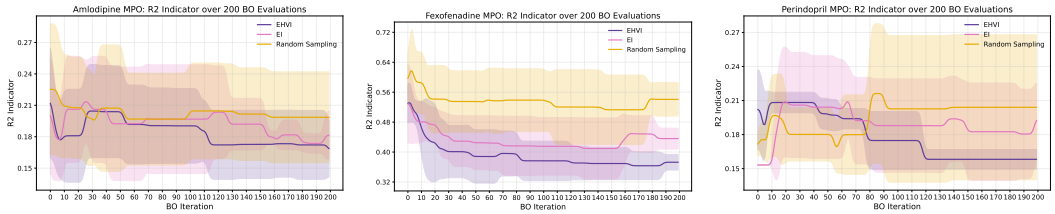

(a) Amlodipine MPO (b) Fexofenadine MPO (c) Perindopril MPO

Figure 1: Hypervolume indicator (HVI) over 200 Bayesian optimization iterations for each MPO task. EHVI consistently achieves higher hypervolume than scalarized EI and random sampling, with faster convergence and greater final front coverage. Shaded areas represent standard deviation over 3 random seeds.

## 4.2 $R^2$ Indicator

In Figure 2 below, EHVI exhibits clear superiority in Pareto front approximation across all tasks. In Fexofenadine, it achieves the lowest and most stable $R^2$ scores, with a wide and persistent gap over scalarized EI. Amlodipine shows modest separation, with EHVI attaining consistently lower scores after early fluctuations. In Perindopril, EHVI dominates from mid-optimization onward, yielding

the most stable and lowest variance estimates. These trends indicate that EHVI not only expands the front efficiently but also produces more uniformly optimal trade-offs across objectives.

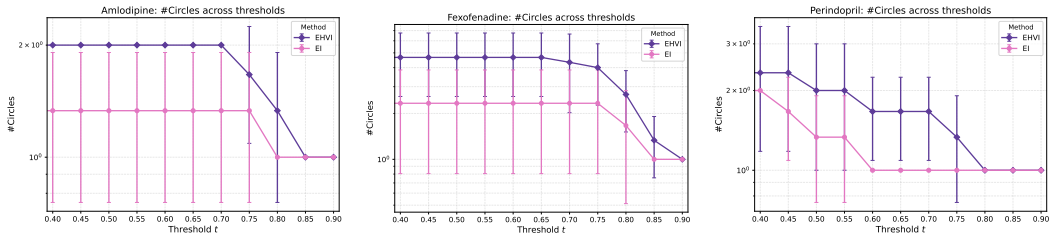



(a) Amlodipine MPO (b) Fexofenadine MPO (c) Perindopril MPO

Figure 2: $R^2$ indicator across 200 Bayesian optimization iterations for each MPO task. Lower values reflect better approximation of the true Pareto front under varying utility directions. EHVI consistently achieves lower $R^2$ values than scalarized EI and random sampling, indicating superior convergence toward the reference front. Shaded regions show standard deviation over 3 random seeds.

### 4.3 #Circles Metric

EHVI demonstrates superior or comparable chemical diversity across all MPO tasks shown in Figure 3. In Fexofenadine, EHVI clearly outpaces EI at thresholds $t \geq 0.60$, uncovering significantly more distinct structural motifs. Perindopril shows a similar advantage, with EHVI sustaining diversity across the full range of thresholds. In Amlodipine, both methods perform similarly at low to mid thresholds, but EHVI maintains higher diversity beyond $t = 0.75$ suggesting enhanced exploration of chemically dissimilar optima. These patterns highlight EHVI's ability to balance objective performance with structural novelty.




(a) Amlodipine MPO (b) Fexofenadine MPO (c) Perindopril MPO

Figure 3: Structural diversity assessed using the #Circles metric across increasing Tanimoto distance thresholds. Higher values indicate broader exploration of structurally distinct regions of the chemical space. EHVI consistently maintains or exceeds the diversity of scalarized EI, particularly at stricter thresholds. Error bars denote standard deviation across 3 random seeds.

## 5 Discussion

This study offers a systematic empirical investigation of Pareto-based Bayesian optimization using Expected Hypervolume Improvement (EHVI), contrasting it against a widely used single-objective acquisition strategy: scalarized Expected Improvement (EI) with fixed weights. We focus on a fixed-weight baseline, commonly used in practical pipelines. This allows us to isolate the benefits of Pareto-aware acquisition in molecular optimization, without confounding from scalarization strategy variation.

Our findings demonstrate that EHVI yields more sample-efficient exploration across a range of realistic multi-property optimization (MPO) tasks. These gains are reflected across three orthogonal metrics—hypervolume coverage, front approximation accuracy ($R^2$), and structural diversity—and are statistically robust across random seeds, as shown in Appendix A.2 (Tables 3, 4). Crucially, these improvements are not due to changes in surrogate fidelity, kernel choice, or representation

capacity: both EHVI and scalarized EI operate under the same model assumptions and input features, underscoring the importance of acquisition strategy alone in driving performance.

Another promising direction involves evaluating the role of Monte Carlo (MC) sampling in EHVI's performance. Our current setup uses 1000 MC samples per candidate evaluation, which strikes a balance between computational overhead and estimate fidelity. However, increasing this number could reduce integration variance and enhance the accuracy of hypervolume estimates, potentially accelerating convergence and improving solution quality. A dedicated ablation study could clarify whether EHVI disproportionately benefits from higher sampling rates—particularly in higher-dimensional tasks where MC estimation errors can compound [31]. Similarly, future work should investigate how the precision of the Monte Carlo estimate—i.e., the amount of stochastic noise in the acquisition value—impacts optimization behavior. In particular, it remains unclear whether EHVI or scalarized EI are more sensitive to noisy acquisition signals, which could affect candidate selection and convergence speed.

We also emphasize the role of molecular representation. This study employed full-dimensional count-based ECFP vectors with a MinMax kernel—an information-rich but high-dimensional descriptor space. While this aligns with recent findings that preserving fingerprint fidelity improves surrogate model accuracy [35], it also introduces challenges related to model scalability and sample efficiency. Benchmarking EHVI and EI using reduced-dimensional or contrastively learned fingerprint embeddings could reveal how robust each method is to representation compression or abstraction.

Despite the promising results presented here, our surrogate models use independent Gaussian Processes without adaptive hyperparameter tuning. This choice prioritizes control and reproducibility, but may limit adaptability to evolving posterior landscapes, particularly in complex or noisy objective settings. Exploring more expressive surrogates—such as deep kernel methods, ensembling, or dynamically updated GPs—could enhance fidelity and robustness. Additionally, while EHVI and scalarized EI offer strong baselines, other acquisition strategies deserve future evaluation. For example, ParEGO [23], PESMO [20], and generative diversity-promoting methods like Multi-Objective GFlowNets [21] offer orthogonal strengths. Expanding evaluations to include noisy, constrained, or synthesis-feasible objectives would further establish the practical utility of Pareto-aware acquisition.

## 6 Conclusion

We conducted a focused comparison between Pareto-based Bayesian optimization (EHVI) and scalarized Expected Improvement (EI) with fixed weights across multi-objective molecular design tasks. EHVI consistently outperforms EI, achieving higher hypervolume, lower $R^2$ values, and greater structural diversity. Statistical effect size analysis further confirms these gains are meaningful despite limited runs. Unlike scalarized EI, which optimizes a fixed utility, EHVI explicitly captures trade-offs, leading to more balanced and diverse Pareto fronts. These results demonstrate the value of Pareto-aware acquisition in data-constrained settings and motivate future work on adaptive surrogates, representation learning, and diversity-aware BO strategies.

## References


[1] Alaleh Ahmadianshalchi, Syrine Belakaria, and Janardhan R. Doppa. Pareto front-diverse batch multi-objective bayesian optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12017–12025, 2024. doi: 10.1609/aaai.v38i10.28951. URL https://ojs.aaai.org/index.php/AAAI/article/view/28951.

[2] V. J. Bowman Junior. On the relationship of the tchebycheff norm and the efficient frontier of multiple-criteria objectives. *Springer Science*, 1976. doi: 10.1007/978-3-642-87563-2_5.

[3] Vlad. M Brochu, Eric Cora and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv*, 2010. URL https://doi.org/10.48550/arXiv.1012.2599.

[4] Nathan Brown, Marco Fiscato, Marwin H.S. Segler, and Alain C. Vaucher. Guacamol: Benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, March 2019. ISSN 1549-960X. doi: 10.1021/acs.jcim.8b00839. URL http://dx.doi.org/10.1021/acs.jcim.8b00839.

[5] Norman Cliff. *Ordinal methods for behavioral data analysis*. Psychology Press, 2014.

[6] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[7] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization, 2020. URL https://arxiv.org/abs/2006.05078.

[8] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6 (2):182–197, 2002.

[9] Matthias Ehrgott. *Multicriteria optimization*, volume 491. Springer Science & Business Media, 2005.

[10] Sean Ekins, J Dana Honeycutt, and James T Metz. Evolving molecules using multi-objective optimization: applying to adme/tox. *Drug Discovery Today*, 15(11-12):451–460, 2010. doi: 10.1016/j.drudis.2010.04.003.

[11] Michael T. M. Emmerich, André H. Deutz, and Jan Willem Klinkenberg. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pages 2147–2154, 2011. doi: 10.1109/CEC.2011.5949880.

[12] C.M. Fonseca, L. Paquete, and M. Lopez-Ibanez. An improved dimension-sweep algorithm for the hypervolume indicator. In *2006 IEEE International Conference on Evolutionary Computation*, pages 1157–1163, 2006. doi: 10.1109/CEC.2006.1688440.

[13] Jenna C Fromer and Connor W Coley. Computer-aided multi-objective optimization in small molecule discovery. *Patterns*, 4(2), 2023. URL https://www.cell.com/patterns/fulltext/S2666-3899(23)00001-6.

[14] Jenna C. Fromer, David E. Graff, and Connor W. Coley. Pareto optimization to accelerate multi-objective virtual screening. *Digital Discovery*, 3:467–481, 2024. doi: 10.1039/D3DD00227F. URL http://dx.doi.org/10.1039/D3DD00227F.

[15] Wenhao Gao, Tianfu Fu, Jimeng Sun, and Connor W. Coley. Sample efficiency matters: A benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. doi: 10.48550/arXiv.2206.12411.

[16] Daniel Golovin and Qiuyi Zhang. Random hypervolume scalarizations for provable multi-objective black box optimization, 2020. URL https://arxiv.org/abs/2006.04655.

[17] Abhijith M. Gopakumar, Prasanna V. Balachandran, Dezhen Xue, James E. Gubernatis, and Turab Lookman. Multi-objective optimization for materials discovery via adaptive design. *Scientific Reports*, 8(1):3738, 2018. doi: 10.1038/s41598-018-21936-3. URL https://doi.org/10.1038/s41598-018-21936-3.

[18] Michael Pilegaard Hansen and Andrzej Jaszkiewicz. Evaluating the quality of approximation to the non-dominated set. *IMM Technical Report*, page 29, 03 1998.

[19] S. Helfrich, A. Herzel, and S. Ruzika. Using scalarizations for the approximation of multiobjective optimization problems: Towards a general theory. *Mathematical Methods of Operations Research*, 100:27–63, 2024. doi: 10.1007/s00186-023-00823-2. URL https://doi.org/10.1007/s00186-023-00823-2.

[20] Daniel. Hernandez-Lobato, Jose Miguel Hernandez-Lobato, Shah Amar, and Adams Ryan. Predictive entropy search for multi-objective bayesian optimization. *Proceedings of The 33rd International Conference on Machine Learning*, 48:1492–1501, 2016.

[21] Moksh Jain, Sharath Chandra Raparthy, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Yoshua Bengio, Santiago Miret, and Emmanuel Bengio. Multi-objective gflownets, 2023. URL https://arxiv.org/abs/2210.12765.

[22] Yeonju Kim, Zherong Pan, and Kris Hauser. Mo-bbo: Multi-objective bilevel bayesian optimization for robot and behavior co-design. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9877–9883. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9561846.

[23] Joshua Knowles. Parego: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation*, 10 (1):50–66, 2006.

[24] Mina Konakovic Lukovic, Yunsheng Tian, and Wojciech Matusik. Diversity-guided multiobjective bayesian optimization with batch evaluations. *Advances in Neural Information Processing Systems*, 33:17708–17720, 2020.

[25] Avan Kumar, Kamal K Pant, Sreedevi Upadhyayula, and Hariprasad Kodamana. Multiobjective bayesian optimization framework for the synthesis of methanol from syngas using interpretable gaussian process models. *ACS omega*, 8(1):410–421, 2022. URL https://doi.org/10.1021/acsomega.2c04919.

[26] Xi Lin, Yilu Liu, XiaoYuan Zhang, Fei Liu, Zhenkun Wang, and Qingfu Zhang. Few for many: Tchebycheff set scalarization for many-objective optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2405.19650.

[27] Tadahiko Murata, Hisao Ishibuchi, and Hideo Tanaka. Multi-objective genetic algorithm and its applications to flowshop scheduling. *Computers & Industrial Engineering*, 30(4):957–968, 1996. doi: 10.1016/0360-8352(96)00045-9. URL https://doi.org/10.1016/0360-8352(96)00045-9.

[28] Christos A Nicolaou and Nathan Brown. Multi-objective optimization methods in drug design. *Drug Discovery Today: Technologies*, 10(3):e427–e435, 2013. doi: 10.1016/j.ddtec.2013.02.001.

[29] Biswajit Paria, Kirthevasan Kandasamy, and Barnabás Póczos. A flexible framework for multiobjective bayesian optimization using random scalarizations, 2019. URL https://arxiv.org/abs/1805.12168.

[30] Ji Won Park, Samuel Stanton, Saeed Saremi, Andrew Watkins, Henri Dwyer, Vladimir Gligorijevic, Richard Bonneau, Stephen Ra, and Kyunghyun Cho. Propertydag: Multi-objective bayesian optimization of partially ordered, mixed-variable properties for biological sequence design. *arXiv preprint arXiv:2210.04096*, 2022. URL https://doi.org/10.48550/arXiv.2210.04096.

[31] G Peter Lepage. A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, 27(2):192–203, 1978. ISSN 0021-9991. doi: https://doi.org/10.1016/0021-9991(78)90004-9. URL https://www.sciencedirect.com/science/article/pii/0021999178900049.

[32] Liva Ralaivola, S. J. Swamidass, H. Saigo, and P. Baldi. Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110, 2005. doi: 10.1016/j.neunet.2005.07.009.

[33] Carl Edward Rasmussen and Christopher K.I Williams. *Gaussian Processes for Machine Learning*. MIT Press Direct, 2005. URL https://doi.org/10.7551/mitpress/3206.001.0001.

[34] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. PMID: 20426451.

[35] Austin Tripp and José Miguel Hernández-Lobato. Diagnosing and fixing common problems in bayesian optimization for molecule design, 2024. URL https://arxiv.org/abs/2406.07709.

[36] Austin Tripp, Gregor N. C. Simm, and Jose Miguel Hernandez-Lobato. A fresh look at de novo molecular design benchmarks. *NeurIPS Workshop 2021 AI4Science*, 2021. URL https://openreview.net/pdf?id=gS3XMun4cl_.

[37] Xing Wang, Bing Wang, Joshua Pinskier, Yue Xie, James Brett, Richard Scalzo, and David Howard. Fin-bayes: A multi-objective bayesian optimization framework for soft robotic fingers. *Soft Robotics*, 11(5):791–801, 2024. doi: https://doi.org/10.1089/soro.2023.0134.

[38] Yutong Xie, Ziqiao Xu, Jiaqi Ma, and Qiaozhu Mei. How much space has been explored? measuring the chemical space covered by databases and machine-generated molecules, 2023. URL https://arxiv.org/abs/2112.12542.

[39] Qingfu Zhang and Hui Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007.

[40] Yiheng Zhu, Jialu Wu, Chaowen Hu, Jiahuan Yan, Tingjun Hou, Jian Wu, et al. Sample-efficient multi-objective molecular optimization with gflownets. *Advances in Neural Information Processing Systems*, 36:79667–79684, 2023.

[41] Eckart Zitzler and Lothar Thiele. Multiobjective optimization using evolutionary algorithms — a comparative case study. In Agoston E. Eiben, Thomas Bäck, Marc Schoenauer, and Hans-Paul Schwefel, editors, *Parallel Problem Solving from Nature — PPSN V*, pages 292–301, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

# A Appendix

## A.1 Final Performance Metrics after 200 BO evaluations

Table 1: Final hypervolume (mean $\pm$ std) after 200 BO evaluations for each MPO task, computed over 3 random seeds.

| Task | EHVI | Scalarized EI |
|---|---|---|
| Fexofenadine | $0.4022 \pm 0.0661$ | $0.3492 \pm 0.0190$ |
| Amlodipine | $0.2421 \pm 0.0425$ | $0.2220 \pm 0.0251$ |
| Perindopril | $0.2080 \pm 0.0016$ | $0.2088 \pm 0.0230$ |

Table 2: Final $R^2$ values (mean $\pm$ std) after 200 BO evaluations for each MPO task, computed over 3 random seeds.

| Task | EHVI | Scalarized EI |
|---|---|---|
| Fexofenadine | $0.3728 \pm 0.0204$ | $0.4360 \pm 0.0293$ |
| Amlodipine | $0.1649 \pm 0.0203$ | $0.1816 \pm 0.0212$ |
| Perindopril | $0.1582 \pm 0.0087$ | $0.1953 \pm 0.0322$ |

### A.2 Cohen's $d$ and Cliff's Delta

Table 3: Effect size metrics comparing EHVI and scalarized EI on the hypervolume indicator across three MPO tasks.

| **Task** | **Cohen's $d$** | **Cliff's Delta** |
|---|---|---|
| Fexofenadine | 1.093 | 0.556 |
| Amlodipine | 0.576 | 0.333 |
| Perindopril | -0.050 | 0.333 |

Table 4: Effect size between EHVI and scalarized EI for $R^2$ values across each task. Negative values indicate better performance by EHVI.

| **Task** | **Cohen's $d$** | **Cliff's Delta** |
|---|---|---|
| Fexofenadine | $-2.560$ | $-1.000$ |
| Amlodipine | $-0.770$ | $-0.556$ |
| Perindopril | $-1.602$ | $-0.778$ |