BOtied: Multi-objective Bayesian optimization with tied multivariate ranks

Anonymous Authors[1]

Abstract

Many biological applications require joint optimization of multiple, potentially competing objectives. Multi-objective Bayesian optimization (MOBO) is a sample-efficient framework for identifying Pareto-optimal solutions. At the heart of MOBO is the acquisition function, which determines the next candidate to evaluate by navigating the best compromises among the objectives. We propose the CDF indicator, a Pareto-compliant metric for evaluating the quality of approximate Pareto sets. and an acquisition function, called BOTIED, based on the CDF indicator. BOTIED can be implemented efficiently with copulas, a statistical tool for modeling complex, high-dimensional distributions. We benchmark BOTIED against common acquisition functions, including EHVI and random scalarization (ParEGO), in a series of synthetic and real-data experiments. BOTIED performs on par with the baselines across datasets and metrics while being computationally efficient.

1. Introduction

Bayesian optimization (BO) has demonstrated promise in a variety of scientific and industrial domains where the goal is to optimize an expensive black-box function using a limited number of potentially noisy function evaluations [25; 7] . In BO, we fit a probabilistic surrogate model on the available observations so far. Based on the model, the acquisition function determines the next candidate to evaluate by balancing exploration (evaluating highly uncertain candidates) with exploitation (evaluating designs believed to maximize the objective). Often, applications call for joint optimization of multiple, potentially competing objectives. Unlike in single-objective settings, a single optimal solution may not exist and we must identify a set of solutions that represents the best compromises among the multiple objectives. The acquisition function in multi-objective Bayesian optimization (MOBO) navigates these trade-offs as it guides the optimization toward regions of interest.

Many MO acquisition functions without scalarization, such as expected hypervolume improvement [EHVI; 12; 13; 8; 9] or entropy search, involve high-dimensional integrals and scale poorly with increasing numbers of objectives. EHVI and random scalarization [20; 24] are sensitive to non-informative transformations of the objectives, such as rescaling of one objective relative to another or monotonic transformations of individual objectives. To address these challenges, we propose BOTIED[1], a novel acquisition function based on multivariate ranks. We show that BOTIED has the desirable property of being invariant to relative rescaling or monotonic transformations of the objectives. While it maintains the multivariate structure of the objective space, its implementation has highly favorable time complexity and we report wall-clock time competitive with scalarization.

In Fig. 1(a), we present the intuition behind multivariate ranks. Consider a maximization setup over two objectives where we seek to identify solutions on the true Pareto frontier (red curves), hypothetical and inaccessible to us. Suppose we have many candidates, represented as circular posterior blobs in the objective space, where the posteriors have been inferred from our probabilistic surrogate model. For simplicity, assume the posterior widths (uncertainties) are comparable among the candidates. Let us consider each candidate individually. How do we estimate each candidate's proximity to the true Pareto frontier? Our surrogate predicts the candidate shaded in blue to have high values in both objectives and, unbeknownst to us, it happens to lie on the true Pareto front. On the other hand, the orange candidate is predicted to be strictly dominated by the blue counterpart. The areas of regions bounded from above by the candidates corroborate this ordering, as shown in the leftmost panel; the hypervolume (HV) dominated by the blue candidate (see Eq. 2.1) is bigger than that of the orange. Alternatively, we can compute multivariate ranks of the candidates (middle panel). Consistent with the HV ordering, the blue candidate is ranked higher, at 1, than the orange candidate, at 3. Note that, due to orthogonality, there may be *ties* among

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

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

[1]The name choice stems from non-dominated candidates considered as "tied".

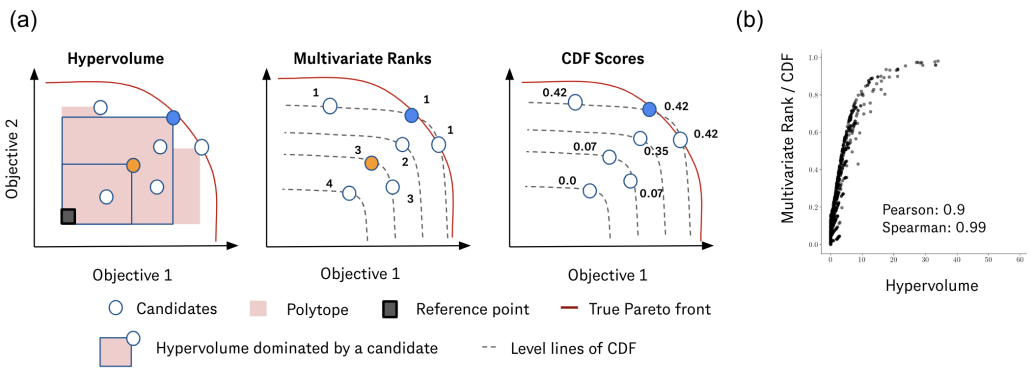

Figure 1: (a) Conceptual summary of BOTIED: Here, the blue candidate is predicted to dominate orange with respect to both objectives. The HV indicator is consistent with this ordering; the area of the box bounded by the blue candidate is bigger than that bounded by the orange. Multivariate ranks and CDF scores, used in BOTIED, are also more favorable for the blue candidate. (b) The CDF scores closely trace HV.

the candidates. Ranking in high dimensions is not a trivial task, as there is no natural ordering in Euclidean spaces when $M \geq 2$. To compute multivariate ranks, we propose to use the (joint) cumulative distribution function (CDF) defined as the probability of a sample having greater function value than other candidates, $F_Y(y) = P(f(X) \leq y)$, where $y = f(x)$. The gray dashed lines indicate the level lines of the CDF. The level line at $F(\cdot) = 1$ is the Pareto frontier estimated by our CDF. As Fig. 1(b) shows, the CDF scores themselves closely trace HV as well.

Motivated by the natural interpretation of multivariate ranks as a multi-objective indicator, we make the following contributions: (i) a new Pareto-compliant performance criterion, the CDF indicator, and (ii) a scalable and robust acquisition function based on the multirank, BOTIED.

2. MOBO with tied multivariate ranks

When there are multiple objectives of interest, a single best design may not exist. Suppose there are M objectives, $f : \mathcal{X} \to \mathbb{R}^M$. The goal of multi-objective BO is to identify the set of *Pareto-optimal* solutions such that improving one objective within the set leads to worsening another. We say that $\boldsymbol{x}$ dominates $\boldsymbol{x}'$, or $f(\boldsymbol{x}) \succ f(\boldsymbol{x}')$, if $f_m(\boldsymbol{x}) \geq f_m(\boldsymbol{x}')$ for all $m \in \{1, \ldots, M\}$ and $f_m(\boldsymbol{x}) > f_m(\boldsymbol{x}')$ for some m. The set of *non-dominated* solutions $\mathscr{X}^*$ is defined in terms of the Pareto frontier (PF) $\mathcal{P}^*$,

$$\mathscr{X}^\star = \{\boldsymbol{x} : f(\boldsymbol{x}) \in \mathcal{P}^\star\},$$

where $\mathcal{P}^\star = \{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X},\ \nexists \boldsymbol{x}' \in \mathcal{X}\ s.t.\ f(\boldsymbol{x}') \succ f(\boldsymbol{x})\}$.

MOBO algorithms typically aim to identify a finite subset of $\mathscr{X}^\star$, which may be infinite, within a reasonable number of iterations. One way to measure the quality of an approximate PF $\mathcal{P}$ is to compute the hypervolume (HV) $\mathrm{HV}(\mathcal{P}|\boldsymbol{r}_{\mathrm{ref}})$ of the polytope bounded from above by $\mathcal{P}$ and from below by $\boldsymbol{r}_{\mathrm{ref}}$, where $\boldsymbol{r}_{\mathrm{ref}} \in \mathbb{R}^M$ is a user-specified *reference point*. More specifically, the HV indicator for a set A is

$$I_{\mathrm{HV}}(A) = \int_{\mathbb{R}^M} \mathbb{I}[\boldsymbol{r}_{\mathrm{ref}} \preceq \boldsymbol{y} \preceq A] d\boldsymbol{y}. \tag{2.1}$$

We obtain the EHVI acquisition function if we take

$$u_{\mathrm{EHVI}}(\boldsymbol{x}, \hat{f}, \mathcal{D}) = \mathrm{HVI}(\mathcal{P}', \mathcal{P}|\boldsymbol{r}_{\mathrm{ref}}) = [I_{\mathrm{HV}}(\mathcal{P}'|\boldsymbol{r}_{\mathrm{ref}}) - I_{\mathrm{HV}}(\mathcal{P}|\boldsymbol{r}_{\mathrm{ref}})]_+, \tag{2.2}$$

where $\mathcal{P}' = \mathcal{P} \cup \{\hat{f}(\boldsymbol{x})\}$ [12; 13].

In MOBO, it is common to evaluate the quality of an approximate Pareto set $\mathcal{X}$ by computing its distance from the optimal Pareto set $\mathcal{X}^*$ in the objective space, or $d(f(\mathcal{X}), f(\mathcal{X}^*))$. The distance metric $d : 2^{\mathcal{Y}} \times 2^{\mathcal{Y}} \to \mathbb{R}$ quantifies the difference between the sets of objectives, where $2^{\mathcal{Y}}$ is the power set of the objective space $\mathcal{Y}$. Existing work in MOBO mainly focuses on the difference in HV, or HVI.

In the following, the *(weak) Pareto-dominance* relation is used as a preference relation $\succcurlyeq$ on the search space X indicating that a solution x is at least as good as a solution $y (x \succcurlyeq y)$ if and only if $\forall 1 \leq i \leq M : f_i(x) \geq f_i(y)$. This relation can be canonically extended to sets of solutions where a set $A \subseteq X$ weakly dominates a set $B \subseteq X (A \succcurlyeq B)$ iff $\forall y \in B\ \exists x \in A : x \succcurlyeq y$ [31]. For MOBO, we need quality indicators that assign each approximation set a real number, i.e., a (unary) indicator I is a function $I : \Omega \to \mathbb{R}$ [31]. One important feature an indicator should have is *Pareto compliance* [14], i.e., it must not contradict the order induced by the Pareto dominance relation, i.e. it should be Pareto-compliant. In particular, this means that whthat the indicator value of A is $A \succcurlyeq B$.

2.1. CDF Indicator

Here we propose the CDF indicator for measuring the quality of Pareto approximations.

Definition 1. *(Cumulative Distribution Function): The cdf of a real-valued random variable Y is the function given by:*

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} p_Y(t)dt. \tag{2.3}$$

i.e. it represents the probability that the r.v. Y takes on a value less than or equal to y.

For more then two variables, the joint CDF is given by:

$$F_{Y_1,\ldots,Y_M} = \int_{(-\infty,\ldots,-\infty)}^{(y_1,\ldots,y_M)} p_Y(\mathbf{t})dt \tag{2.4}$$

Definition 2. *(CDF Indicator). The CDF indicator I_F is defined as the maximum multivariate rank*

$$I_{F_\mathbf{Y}}(A) := \max_{y \in A} F_\mathbf{Y}(y) \tag{2.5}$$

where A is an approximation set in Ω.

Next we show that this indicator is compliant with the concept of Pareto dominance.

Theorem 1. *For any arbitrary approximation sets $A \in \Omega$ and $B \in \Omega$, it holds*

$$A \succcurlyeq B \wedge B \not\succcurlyeq A \Rightarrow I_F(A) \geq I_F(B). \tag{2.6}$$

The proof can be found in Appendix B. We illustrate the robust properties of the CDF indicator in Fig. 2.

Estimation of the CDF indicator Estimating the multivariate joint distribution $F_\mathbf{Y}$ is a challenging task. A naive approach would involve estimating the multivariate density function and then computing the integral, which is computationally intensive. We thus turn to *copulas* [23; 3], statistical tool for flexible density estimation in higher dimensions.

Theorem 2 (Sklar's theorem [26])**.** *The continuous random vector $Y = (Y_1, \ldots, Y_M)$ has joint a distribution F and marginal distributions $Y_1, \ldots, F_M$ if and only if there exist a unique copula C, which is the joint distribution of $U = (U_1, \ldots, U_M) = F_1(Y_1), \ldots, F_d(Y_M)$.*

From Sklars' theorem, we note that a copula is a multivariate distribution function $C : [0,1]^M \to [0,1]$ that joins (i.e. couples) uniform marginal distributions:

$$F(y_1, \ldots, y_M) = C(F_1(y_1), \ldots, F_d(y_M)). \tag{2.7}$$

By computing the copula function, we also obtain access to the multivariate CDF, and by construction to the multivariate ranking. The benefits of using copulas as estimators for the CDF indicator are three fold: (i) Scalability and flexible estimation in higher dimensional objective spaces, (ii) Scale invariance wrt different objectives (Fig. 2), (iii) Invariance Under Monotonic Transformations of the objectives.

Vine copulas for high-dimensional CDFs A copula can be modeled following a parametric family depending on the shape of the dependence structure, such as Clayton copula with lower tail dependence or Gumbel copula with upper tail. For additional flexibility as well as scalability, [4] has proposed *vine copulas*, a pair-copula construction that allows the factorization of any joint distribution into bivariate copulas. Thus, the copula estimation problem partitions into first determining a graphical model, structure called *vine* consisted of $\frac{M(M-1)}{2}$ trees. Each edge in a tree represents a bivariate copula for which we can choose a parametric or nonparametric estimator. [1] propose efficient algorithms to organize the pair constructions. See Appendix F for an example decomposition using domain knowledge.

2.2. CDF-based acquisition function: BOtied

Suppose we fit a CDF on $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(N_t)}$, the N_t measurements acquired so far. Denote the resulting CDF as $\hat{F}(\cdot; \mathcal{D}_t)$, where we have made explicit the dependence on the dataset up to time t. The BOTIED utility function is as follows:

$$u(\mathbf{x}, \hat{f}, \mathcal{D}_t) = \hat{F}(\hat{f}(\mathbf{x}); \mathcal{D}_t) \tag{2.8}$$

3. Empirical results

Metrics We use the HV indicator, a standard evaluation metric for MOBO, as well as our CDF indicator. We rely on efficient algorithms for HV computation based on hyper-cell decomposition [15; 19] and implemented in `BoTorch` [2].

Baselines We assume noisy function evaluations, so implement noisy versions of all the acquisition functions as baselines. The baseline acquisitions include *NEHVI* (noisy EHVI) [8] *NParEGO* (noisy ParEGO) [20] which uses random augmented Chebyshev scalarization and noisy expected improvement; and *random*. For BOTIED we have implementations v1 and v2, with the only difference being the way of incorporating the variance from the Monte Carlo (MC) predictive posterior samples, either fitting the copula on all of them (v1) or on the means (v2). See Appendix C for algorithms of both versions.

Datasets A toy Penicillin test function [22] ($d = 7$, $M = 3$) simulates the penicillin yield, time to production, and undesired byproduct for various parameters of the production process. This task allows for direct evaluation of f. To emulate a real-world drug design setup, we modify the permeability dataset Caco-2 [29] from the Therapeutics Data Commons database [17; 18]. Permeability is a key property in the absorption, distribution, metabolism, and excretion (ADME) profile of drugs. The Caco-2 dataset consists of 906 drug molecules annotated with experimentally measured rates of passing through a human colon epithelial cancer cell line. We represent each molecule as a concatenation of fingerprint and fragment feature vectors, known as fragprints [28]. The dataset is augmented with five additional properties using RDKit [21], including the drug-likeness score QED [5; 30] and topological polar surface area (TPSA) and refer to the resulting $M = 6$ dataset as Caco-2+. In many cases, subsets

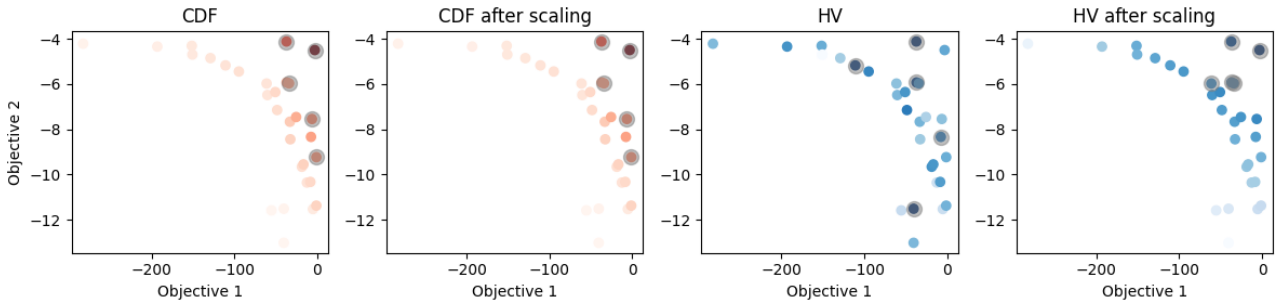

Figure 2: The CDF indicator is more robust to arbitrary rescaling of the objectives than the HV indicator. The color gradient corresponds to value of the indicator for each data point. Gray circles are overlaid on the five selected points with the top indicator scores, where the selection is done in a greedy sequential manner. HV is sensitive to the scales of the objectives.

Table 1: Mean and st dev deviation for HV, computed in the original units, and CDF indicators across synthetic datasets. Higher is better and best per column is bolded.

	Caco2+ (M=3)		Penicillin (M=3)	
	CDF	**HV**	**CDF**	**HV**
BOTIED V1	0.58 (0.06)	11645.63 (629.0)	0.48 (0.02)	319688.6 (17906.2)
BOTIED V2	**0.60 (0.06)**	11208.57(882.21)	**0.49 (0.02)**	318687.7 (17906.2)
NPARETO	0.56 (0.05)	12716.2 (670.12)	0.28 (0.09)	**332203.6 (15701.5)**
NEHVI	0.54 (0.06)	**13224.7 (274.6)**	0.24 (0.05)	318748.9 (2868.64)
RANDOM	0.57 (0.07)	11425.6 (882.4)	0.32 (0.02)	327327.9 (17036)

of these properties (e.g., permeability and TPSA) will be inversely correlated and thus compete with one another during optimization. In late-state drug optimization, the trade-offs become more dramatic and as more properties are added [27]. Demonstrating effective sampling of Pareto-optimal solutions in this setting is thus of great value.

Results and discussion Although there is no single best method across all the datasets, the best numbers are consistently achieved by either BOtied v1 or v2 with NParEGO being a close competitor. In addition to being on par with commonly used acquisition functions, BOTIED is significantly faster than NEHVI as we show in Fig. D. There are two main benefits to using the CDF metric rather than HV for evaluation. First, the CDF is bounded between 0 and 1, with scores close to 1 corresponding to the discovered solutions closest to our approximate Pareto front.[2] Unlike with HV, for which the scales do not carry information about internal ordering, the CDF values have an interpretable scale. Second, applying the CDF metric for different tasks (datasets), we can easily assess how the acquisition performance varies with the specifics of the data. More results on the robustness of the CDF metric and customized small-molecule vine copula estimation can be found in Appendix F.

[2] [6] shows that the zero level lines $F(\cdot) = 0$ correspond to the estimated Pareto front in a minimization setting, equivalent to the one level lines $F(\cdot) = 1$ in the maximization setting.

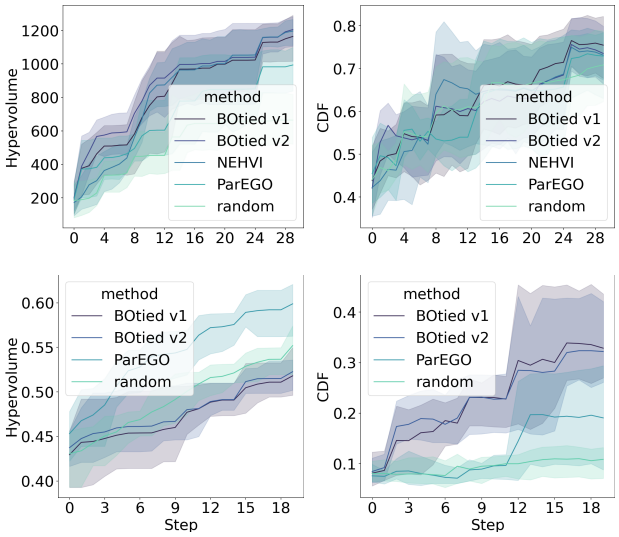

Figure 3: HV/CDF over simulated BO iterations for Branin Currin ($d = 2, M = 2$) [8] and DTLZ ($d = 9, M = 6$) [11]

4. Conclusion

We introduce a new perspective on MOBO by leveraging multivariate ranks computed with CDF scores. We propose a new Pareto-compliant CDF indicator with an efficient implementation using copulas as well as a CDF-based acquisition function. In real and simulated chemical datasets, we have demonstrated our CDF-based estimation of the non-dominated regions allows for greater flexibility, robustness, and scalability compared to existing acquisition functions. This method is general and lends itself to a number of immediate extensions. First, we can encode dependencies between objectives, estimated from domain knowledge, into the graphical vine model. Second, we can accommodate discrete-valued objectives. Finally, whereas we have focused on selecting candidates from a fixed library, the computation of our acquisition function is differentiable and admits gradient-based sampling from the input space.

References

[1] Aas, K. Pair-copula constructions for financial applications: A review. *Econometrics*, 4(4):43, October 2016.

[2] Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

[3] Bedford, T. and Cooke, R. M. Vines – A New Graphical Model for Dependent Random Variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

[4] Bedford, T. and Cooke, R. M. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

[5] 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.

[6] Binois, M., Rullière, D., and Roustant, O. On the estimation of pareto fronts from the point of view of copula theory. *Information Sciences*, 324:270–285, 2015.

[7] Calandra, R., Seyfarth, A., Peters, J., and Deisenroth, M. P. Bayesian optimization for learning gaits under uncertainty: An experimental comparison on a dynamic bipedal walker. *Annals of Mathematics and Artificial Intelligence*, 76:5–23, 2016.

[8] Daulton, S., Balandat, M., and Bakshy, E. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *Advances in Neural Information Processing Systems*, 33:9851–9864, 2020.

[9] Daulton, S., Balandat, M., and Bakshy, E. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *Advances in Neural Information Processing Systems*, 34:2187–2200, 2021.

[10] Daulton, S., Cakmak, S., Balandat, M., Osborne, M. A., Zhou, E., and Bakshy, E. Robust multi-objective bayesian optimization under input noise. *arXiv preprint arXiv:2202.07549*, 2022.

[11] Deb, K. and Gupta, H. Searching for robust pareto-optimal solutions in multi-objective optimization. In *Evolutionary Multi-Criterion Optimization: Third International Conference, EMO 2005, Guanajuato, Mexico, March 9-11, 2005. Proceedings 3*, pp. 150–164. Springer, 2005.

[12] Emmerich, M. *Single-and multi-objective evolutionary design optimization assisted by gaussian random field metamodels*. PhD thesis, Dortmund, Univ., Diss., 2005, 2005.

[13] Emmerich, M. T., Deutz, A. H., and Klinkenberg, J. W. Hypervolume-based expected improvement: Monotonicity properties and exact computation. In *2011 IEEE Congress of Evolutionary Computation (CEC)*, pp. 2147–2154. IEEE, 2011.

[14] Fonseca, C. M., Knowles, J. D., Thiele, L., Zitzler, E., et al. A tutorial on the performance assessment of stochastic multiobjective optimizers. In *Third international conference on evolutionary multi-criterion optimization (EMO 2005)*, volume 216, pp. 240, 2005.

[15] Fonseca, C. M., Paquete, L., and López-Ibánez, M. An improved dimension-sweep algorithm for the hypervolume indicator. In *2006 IEEE international conference on evolutionary computation*, pp. 1157–1163. IEEE, 2006.

[16] Haugh, M. An introduction to copulas. quantitative risk management, 2016.

[17] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*, 2021.

[18] Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*, 2022.

[19] Ishibuchi, H., Akedo, N., and Nojima, Y. A many-objective test problem for visually examining diversity maintenance behavior in a decision space. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, pp. 649–656, 2011.

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

[21] Landrum, G., Tosco, P., Kelley, B., Ric, sriniker, Cosgrove, D., gedeck, Vianello, R., NadineSchneider, Kawashima, E., N, D., Jones, G., Dalke, A., Cole, B., Swain, M., Turk, S., AlexanderSavelyev,

Vaucher, A., Wójcikowski, M., Take, I., Probst, D., Ujihara, K., Scalfani, V. F., guillaume godin, Pahl, A., Berenger, F., JLVarjo, Walker, R., jasondbiggs, and strets123. rdkit/rdkit: 2023_03_1 (q1 2023) release, April 2023. URL https://doi.org/10.5281/zenodo.7880616.

[22] Liang, Q. and Lai, L. Scalable bayesian optimization accelerates process optimization of penicillin production. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[23] Nelsen, R. B. *An introduction to copulas*. Springer Science & Business Media, 2007.

[24] Paria, B., Kandasamy, K., and Póczos, B. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2020.

[25] Romero, P. A., Krause, A., and Arnold, F. H. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

[26] Sklar, A. Fonctions de Répartition à n Dimensions et Leurs Marges. *Publications de L'Institut de Statistique de L'Université de Paris*, 8:229–231, 1959.

[27] Sun, D., Gao, W., Hu, H., and Zhou, S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, 2022.

[28] Thawani, A. R., Griffiths, R.-R., Jamasb, A., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., and Lee, A. A. The photoswitch dataset: A molecular machine learning benchmark for the advancement of synthetic chemistry. *arXiv preprint arXiv:2008.03226*, 2020.

[29] Wang, N.-N., Dong, J., Deng, Y.-H., Zhu, M.-F., Wen, M., Yao, Z.-J., Lu, A.-P., Wang, J.-B., and Cao, D.-S. Adme properties evaluation in drug discovery: prediction of caco-2 cell permeability using a combination of nsga-ii and boosting. *Journal of Chemical Information and Modeling*, 56(4):763–773, 2016.

[30] Wildman, S. A. and Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

[31] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C. M., and Da Fonseca, V. G. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on evolutionary computation*, 7(2): 117–132, 2003.

A. Appendix

B. CDF indicator

Proof: If we have $A \succcurlyeq B \wedge B \not\succcurlyeq A$ then the following two conditions hold: $\forall y \in B \quad \exists x \in A : \mathbf{x} \succcurlyeq \mathbf{y}$ and $\exists \mathbf{x} \in A \quad s.t. \quad \nexists \mathbf{y} \in B : \mathbf{y} \succcurlyeq \mathbf{x}$. Remember that the weak Pareto dominance $\mathbf{x} \succcurlyeq \mathbf{y}$ implies that $\forall 1 \leq i \leq K : f_i(x) \geq f_i(y)$. Every point in the objective space, that is weakly dominated by some element in B, is also weakly dominated by some element in A. From the definition and fundamental property of the CDF, being a monotonic non-decreasing function, it follows that if $\forall 1 \leq i \leq k : f_i(x) \geq f_i(y) \Rightarrow F_{\mathbf{Y}}(x) \geq F_{\mathbf{Y}}(y)$. It is easy to see that by choosing the datapoint with maximum CDF score per set, the set contains the non-dominated solution, will have higher value for the indicator. Since set A contains the non-dominated solution, $I_{F_{\mathbf{Y}}}(A) = \{a_{max} | F_{\mathbf{Y}}(a_{max}) \geq F_{\mathbf{Y}}(a_i), \forall i, 0 \leq i \leq |A|\}$ cannot be worse than the indicator value of $I_{F_{\mathbf{Y}}}(B) = \{b_{max} | F_{\mathbf{Y}}(b_{max}) \geq F_{\mathbf{Y}}(b_j), \forall j, 0 \leq j \leq |B|\}$ and therefore $I_{F_{\mathbf{Y}}}$ is Pareto compliant.

In the low-data regime, empirical Pareto frontiers tend to be noisy. When we have access to domain knowledge about the objectives, we can use it to construct a model-based Pareto frontier using vine copulas. This section describes how to incorporate (1) the known correlations among the objectives to specify the tree structure (vine) and (2) the pairwise joint distributions (including the tail behavior), approximately estimated from domain knowledge, to specify the copula models. The advantages of integrating copula-based estimators for our metric and acquisition function are three fold: (i) scalability — from the convenient pair copula construction of vines, (ii) robustness wrt marginal scales and transformations —, and (iii) domain-aware copula structures — from the explicit encoding of dependencies in the vine copula matrix, including choice of dependence type, e.g., low, high tail dependence.

C. Algorithm: Multi-objective BO with BOtied

Algorithm 1 MOBO with BOTIED, a CDF-based acquisition function

Input: Probabilistic surrogate $\hat{f}$, initial data $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{N_0}$, $\mathcal{X} \subset \mathbb{R}^d$, $\mathbb{R}^M$
Output: Optimal selected subset $\mathcal{D}_T$.
Fit the initial surrogate model $\hat{f}(x_i)$ on $\mathcal{D}_0$.
for $\{t = 1, \ldots, T\}$ **do**
  Sample the candidate pool $x_1, \cdots, x_N \in \mathcal{X}$
  for $\{i = 1, \ldots, N\}$ **do**
    Evaluate the surrogate model $\hat{f}$ on the candidate pool to obtain the posterior $p(\hat{f}(x_i)|\mathcal{D}_{t-1})$
    Draw L predictive samples $f_i^{(j)} \sim p(\hat{f}(x_i)|\mathcal{D}_{t-1})$, for $j \in [L]$
  end for
  Obtain uniform marginals $u_i^{(j)}$ from the pooled posterior samples $f_i^{(j)}$
  $U \leftarrow \{u_i^{(j)}\}_{i \in [N], j \in [L]}$
  Version 1: Fit a vine copula on the uniform marginals on the sample level $C \leftarrow \hat{C}(U)$,
  Version 2: Fit a vine copula on the mean-aggregated uniform marginals $C \leftarrow \hat{C}(\frac{1}{L}\sum_{j=1}^{L} u_i^{(j)})$
  for $\{i = 1, \ldots, N\}$ **do**
    Version 1: Compute the expected CDF score $\mathcal{S}(\boldsymbol{x}_i) = \frac{1}{L}\sum_{j=1}^{L} C(u_i^{(j)})$
    Version 2: Compute the CDF score of the mean ranks $\mathcal{S}(\boldsymbol{x}_i) = C(\frac{1}{L}\sum_{j=1}^{L} u_i^{(j)})$
  end for
  $i' \leftarrow \arg\max_{i \in [N]} \mathcal{S}(\boldsymbol{x}_i)$
  $\mathcal{D}_t \leftarrow \mathcal{D}_{t-1} \cup \{(\boldsymbol{x}_{i'}, \boldsymbol{y}_{i'})\}$
end for$\mathcal{D}_T$

D. MOBO toy experiments

As a numerical testbed, we begin with toy test functions commonly used as BO benchmarks: Branin-Currin [10] ($d = 2$, $M = 2$) and DTLZ [11] ($d = 9$, $M \in \{4, 6, 8\}$).

E. Benefits of using CDF indicator and Vine copulas for estimation

It is important to note that, to be able to estimate a copula, we need to transform the variables of interest to uniform marginals. We do so, by the so-called *probability integral transform (PIT)* of the marginals.

Definition 3. *Probability Integral Transform (PIT) of a random variable Y with distribution F_Y is the random variable $U = F_Y(y)$, which is uniformly distributed $U \sim \mathcal{U}[0, 1]$.*

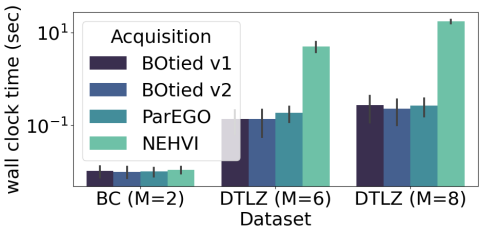

Figure 4: Wall clock time per single call of acquisition function.

The benefit of using copulas as estimators for the CDF indicator are three fold: (i) Scalability and flexible estimation in higher dimensional objective spaces, (ii) Scale invariance wrt different objectives, (iii) Invariance Under Monotonic Transformations of the objectives. These three properties suggest that our indicator is more robust than the widely used Hypervolume indicator, as we will empirically show in the following section. Sklar's theorem, namely the requirement of uniform marginals, immediately implies the following corollary which characterizes the invariance of the CDF indicator to different scales.

Corollary 1. *(Scale Invariance) A copula based estimator for the CDF indicator is scale invariant.*

Corollary 2. *(Invariance Under Monotonic Transformations) Let Y_1, Y_2 be continuous random variables with copula C_{Y_1,Y_2}. If $\alpha, \beta : \mathbb{R} \to \mathbb{R}$ are strictly increasing functions, then:*

$$C_{\alpha(Y_1),\beta(Y_2)} = C_{Y_1,Y_2} \tag{E.1}$$

where $C_{\alpha(Y_1),\beta(Y_2)}$ is the copula function corresponding to variables $\alpha(Y_1)$ and $\beta(Y_2)$.

Corollary 1 follows from the PIT transformation required for copula estimation. The proof for invariance under monotonic transformations based on [16] can be found in Appendix B and without loss of generality can be extended to more than two dimensions. We empirically validate the robustness properties of the copula-based estimator in Fig. 2.

F. Vine Copulas encoding domain knowledge in small molecule BO

Fig. 5 illustrates the use of copulas in the context of optimizing multiple objectives in drug discovery, where data tends to be sparse. In panel (a) we see that, thanks to the separate estimation of marginals and dependence structure, different marginal distributions have the same Pareto front in the PIT space, in which we evaluate our CDF scores. Hence, with copula based estimators, we can guarantee robustness without any overhead for scalarization or standardization of the data as required by counterparts. In panel (b) we show how we can encode domain knowledge of the interplay between different molecular properties in the Caco2+ dataset. Namely, permeability is often highly correlated with ClogP and TPSA, with positive and negative correlation, respectively, which is even more notable at the tails of the data (see panel (a)). Such dependence can be encoded in the vine copula structure and in the choice of copula family for each pair. For example, a rotated Clayton copula was imposed so that the tail dependence between TPSA and Permeability is preserved.

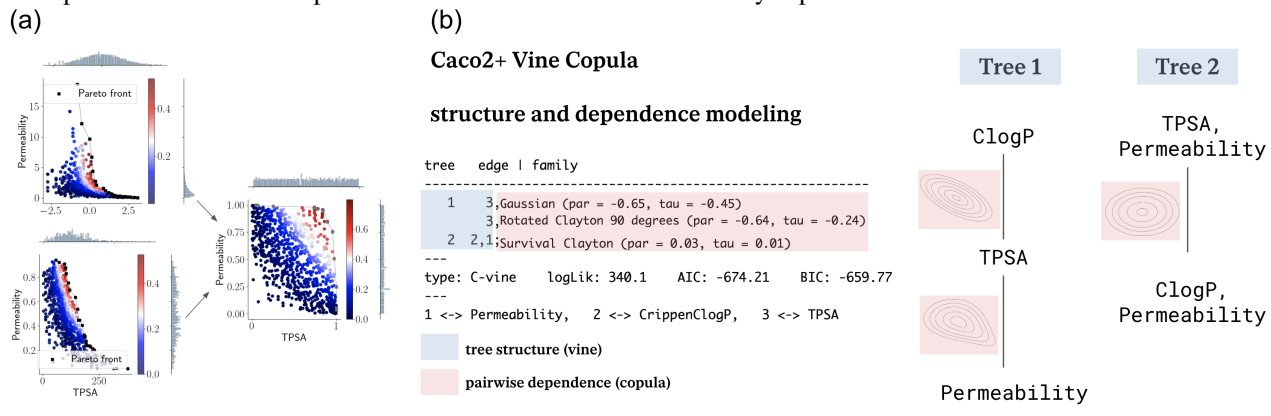

Figure 5: (a). Regardless of the distributions of the marginals, the CDF score from a copula is the same. (b) An example of explicitly encoding domain knowledge in a BO procedure by imposing the in blue tree structure (specifying the matrix representation of the vine) and in pink selection of pairwise dependencies (choice of parametric/nonparametric family).