

Multivariate tails for active molecular design

Ji Won Park

Principal Machine Learning Scientist
Genentech

April 15, 2025
MMLI Symposium



Prescient
Design
A Genentech Accelerator

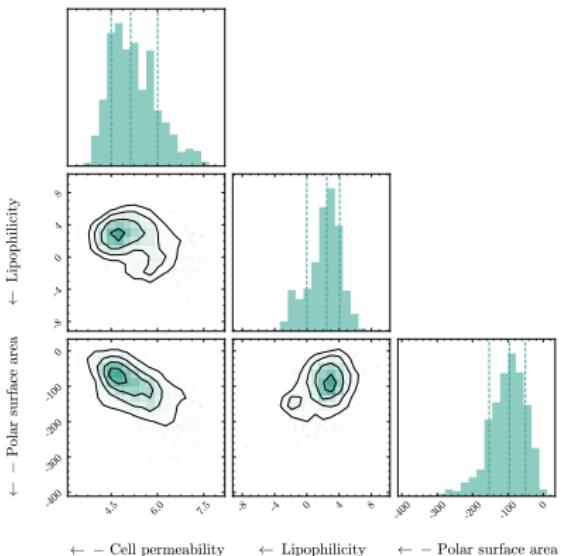
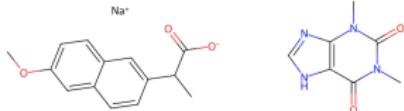
Genentech
A Member of the Roche Group

Based on:

- ▶ **Park, J.W.***, Tagasovska, N.*, Maser, M., Ra, S., and Cho, K. “BOtied: Multi-objective Bayesian optimization with tied multivariate ranks.” ICML (2024). arXiv: [2306.00344](https://arxiv.org/abs/2306.00344)
- ▶ **Park, J.W.**, Tibshirani, R., and Cho, K. “Semiparametric conformal prediction.” AISTATS (2025). arXiv: [2411.02114](https://arxiv.org/abs/2411.02114)

Molecular design: a tale of correlated tails

- ▶ Goal: jointly optimize molecule for multiple competing properties
 - ▶ Molecular properties tend to have long tails¹ and tail correlations²
 - ▶ LLM training and sampling are optimized for average-case behavior



¹Jain et al., "Biophysical properties of the clinical-stage antibody landscape" (2017).

²Wang et al., "ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting" (2016).

Multi-objective optimization

Problem : $\min_{x \in \mathcal{X}} \overbrace{[f_1(x), \dots, f_M(x)]^T}^{f(x)}$



3

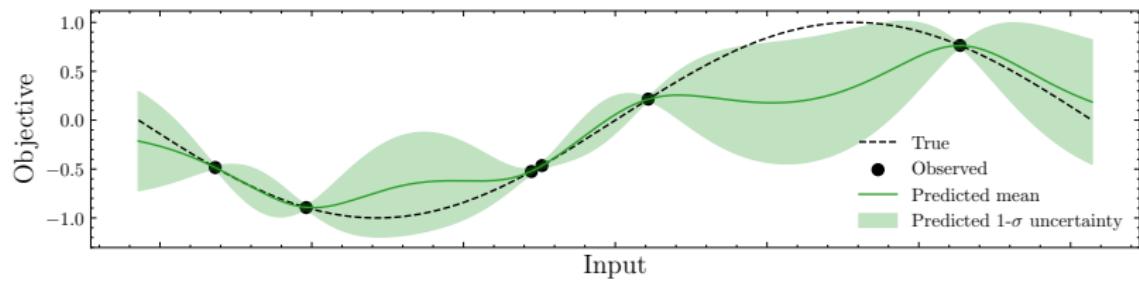
When f is an expensive black-box function (e.g., wet lab protocol), Bayesian optimization offers a sample-efficient method.

³Konakovic Lukovic, Tian, and Matusik, "Diversity-guided multi-objective bayesian optimization with batch evaluations" (2020).

Multi-objective Bayesian optimization (MOBO)

Specify a **probabilistic surrogate model** \hat{f} approximating f .

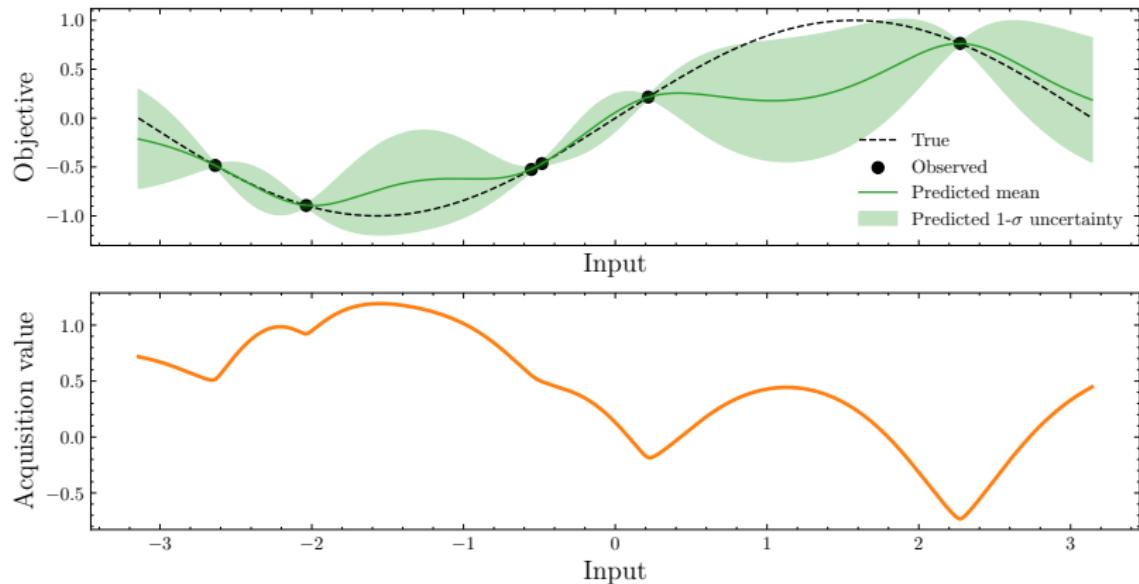
Example: $\hat{f} \sim \mathcal{GP}$ where the spread of $p(\hat{f} | \mathcal{D})$ captures the uncertainty



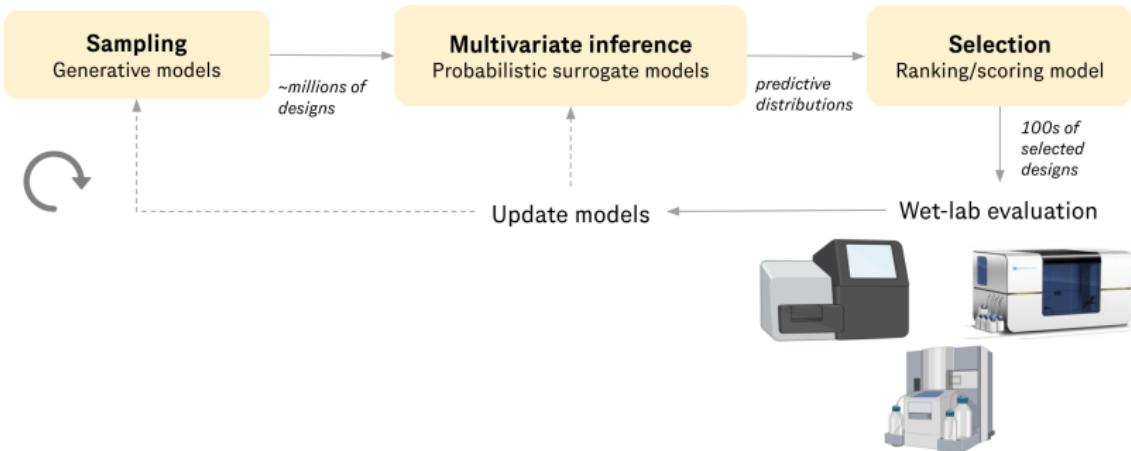
Acquisition function as the decision-making engine

Acquisition function $a^{\hat{f}} : \mathcal{X} \rightarrow \mathbb{R}$ scores each design with predicted “usefulness,” to determine which design to measure next.

- ▶ exploration (of highly uncertain designs)
- ▶ exploitation (of designs believed to be optimal)



Lab-in-the-loop molecular design



1. Fitting the surrogate on $\mathcal{D} = \{(x^{(i)}, f(x^{(i)}))\}_{i=1}^N$, to obtain $p(\hat{f}|\mathcal{D})$
2. Optimizing to obtain $x^* = \operatorname{argmax}_{x \in \mathcal{X}} \hat{a}(\hat{f})(x)$
3. Appending the resulting measurement: $\mathcal{D} \leftarrow \mathcal{D} \cup \{(x^*, f(x^*))\}$

Dominance operators: notation

How to compare vectors in Euclidean spaces when $M > 1$?

Assume minimization. For $y = (y_1, \dots, y_M), z = (z_1, \dots, z_M) \in \mathbb{R}^M$,

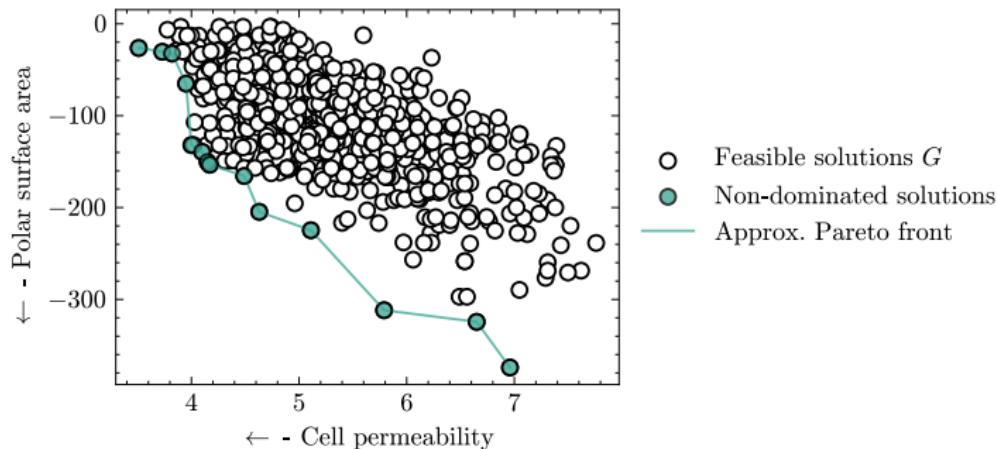
- ▶ “ z **weakly** dominates y ” $z \preccurlyeq y$
 $\iff z_i \leq y_i \ i = 1, \dots, M$
- ▶ “ z **strictly** dominates y ” $z \preceq y$
 $\iff z_i \leq y_i \ \forall i = 1, \dots, M$ and $\exists k : z_k < y_k$
 $\iff z \preccurlyeq y$ and $z \neq y$



Pareto front

For $M > 1$, a single optimal design may not exist.

Pareto front \mathcal{P} is a collection of solutions that are not strictly dominated.



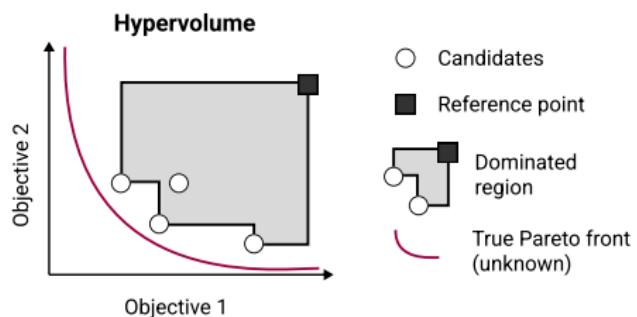
MOBO aims to obtain a **finite approximation** $\hat{\mathcal{P}}$ to the true Pareto front \mathcal{P} .

Quality indicators

Quality indicator $I : 2^{\mathcal{Y}} \rightarrow \mathbb{R}$
evaluates the quality of approximation set $\hat{\mathcal{P}}$.

Hypervolume indicator

Example: **hypervolume** (HV)⁴ of polytope dominated by $\hat{\mathcal{P}}$ and bounded from above by a reference point



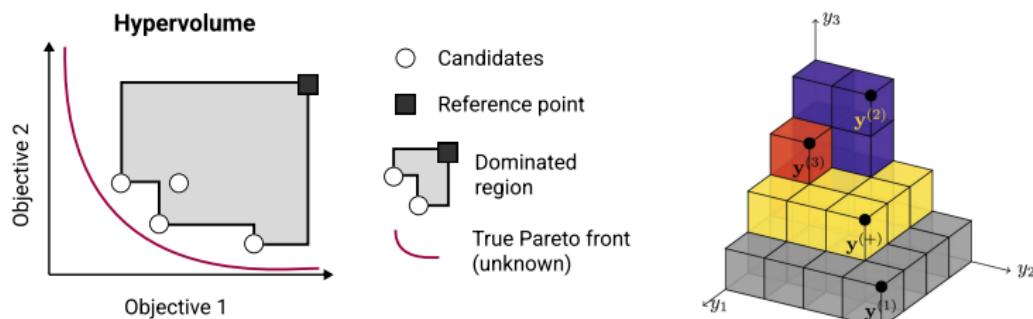
⁴Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

⁵Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Hypervolume indicator: limitations

Example: **hypervolume** (HV)⁴ of polytope dominated by $\hat{\mathcal{P}}$ and bounded from above by a reference point

- $\text{HV} \sim \mathcal{O}(n^{\lfloor \frac{M}{2} \rfloor}) \rightarrow \text{impractical for } M > 4$ despite box decomposition⁵



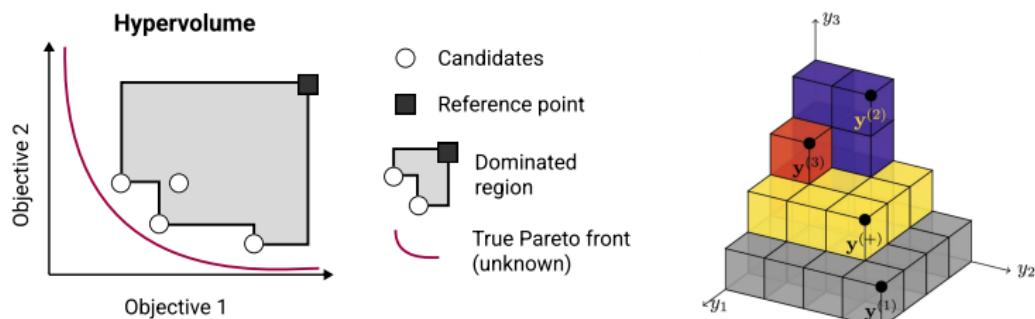
⁴Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

⁵Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Hypervolume indicator: limitations

Example: **hypervolume** (HV)⁴ of polytope dominated by $\hat{\mathcal{P}}$ and bounded from above by a reference point

- ▶ HV $\sim \mathcal{O}(n^{\lfloor \frac{M}{2} \rfloor}) \rightarrow$ impractical for $M > 4$ despite box decomposition⁵
- ▶ Sensitive to rescaling of the objectives, with different natural units



⁴Emmerich, Deutz, and Klinkenberg, "Hypervolume-based expected improvement: Monotonicity properties and exact computation" (2011).

⁵Yang et al., "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization" (2019).

Content

Motivation and Background

- ▶ Drug design: jointly optimizing multiple (tailed) molecular properties
- ▶ A quick primer on multi-objective Bayesian optimization (MOBO)
 - ▶ Quality indicator $I : 2^{\mathcal{Y}} \rightarrow \mathbb{R}$
 - ▶ Acquisition function $a^f : \mathcal{X} \rightarrow \mathbb{R}$

Method

- ▶ Connection between the CDF ranks and the Pareto front
- ▶ BOtied: MOBO based on the CDF

Empirical results

Probabilistic perspective

View molecules as random vectors X .

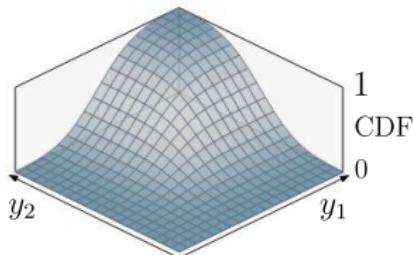
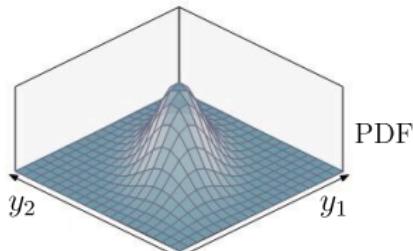
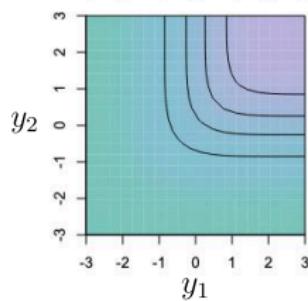
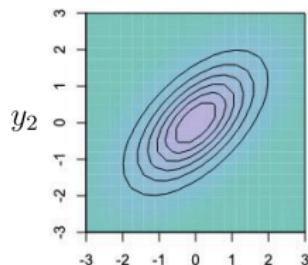
Let $Y = f(X)$, and consider the CDF of Y , F_Y .

Probabilistic perspective

View molecules as random vectors X .

Let $Y = f(X)$, and consider the CDF of Y , F_Y .

$$F_{Y_1, \dots, Y_M}(y) = \int_{(-\infty, \dots, -\infty)}^{(y_1, \dots, y_M)} f_Y(s) ds = \mathbb{P}[Y_1 \leq y_1, \dots, Y_M \leq y_M]$$



Connection between the CDF and the Pareto front

Taking “horizontal slices” at $\alpha \in [0, 1]$ gives the α level line of F_Y ,
 $\partial \mathcal{L}_\alpha^F = \{y' \in G, F_Y(y') = \alpha\}$.

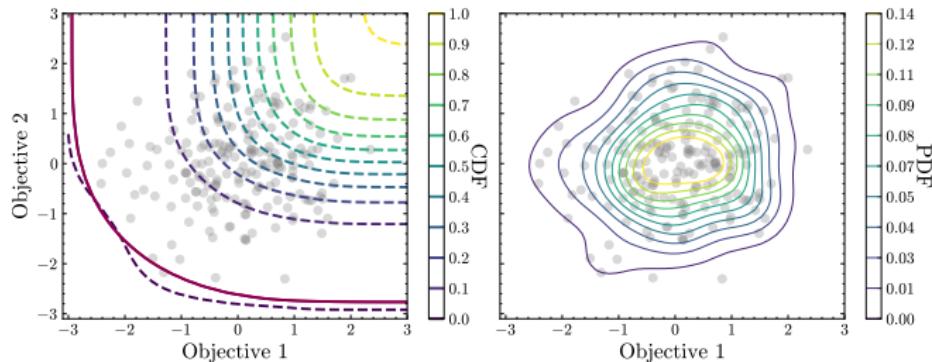
The Pareto front belongs to the zero ($\alpha = 0$) level line of F_Y .⁶

⁶Binois, Rullière, and Roustant, “On the estimation of Pareto fronts from the point of view of copula theory” (2015).

Connection between the CDF and the Pareto front

Taking “horizontal slices” at $\alpha \in [0, 1]$ gives the α level line of F_Y ,
 $\partial\mathcal{L}_\alpha^F = \{y' \in G, F_Y(y') = \alpha\}$.

The Pareto front belongs to the zero ($\alpha = 0$) level line of F_Y .⁶



⁶Binois, Rullière, and Roustant, “On the estimation of Pareto fronts from the point of view of copula theory” (2015).

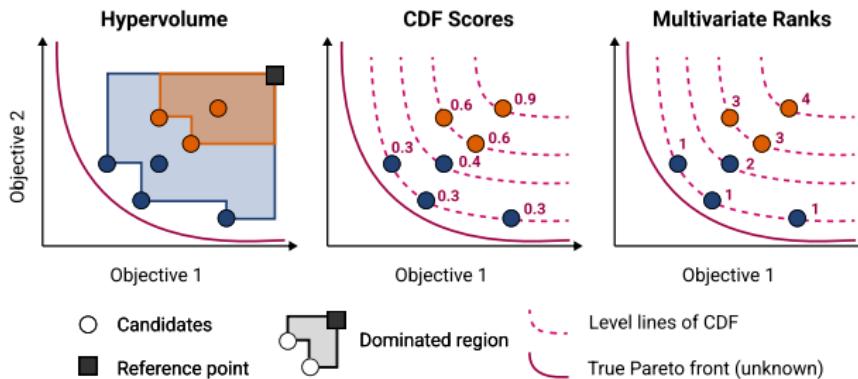
Enter the CDF indicator

We propose $I_{CDF}(A) := \min_{y \in A} F_Y(y)$.

Weak Pareto compliance (Theorem 4.1)

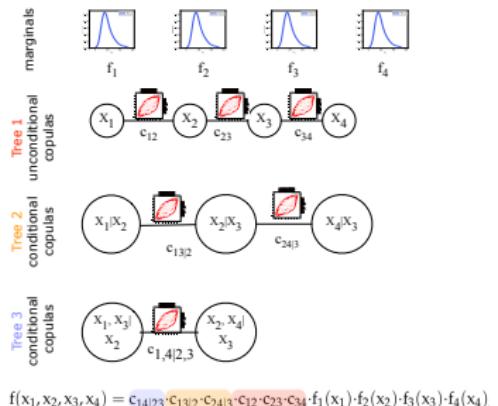
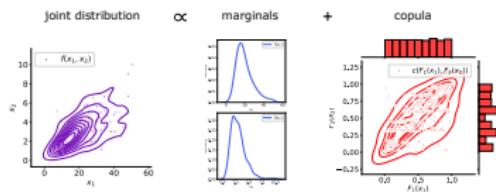
For two approximation sets A and B ,

$$A \leq B \implies I_{CDF}(A) \leq I_{CDF}(B).$$



Efficient fitting of CDF with vine copulas

We can **pairwise decompose** an M -dim copula density into a product of $M(M-1)/2$ bivariate conditional densities (“pair copulas”) organized in a sequence of trees (“vine”)⁷ $\sim \mathcal{O}(nML)$, where $L \in \{1, \dots, M\}$ is depth.

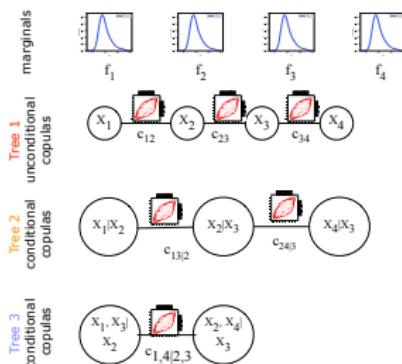
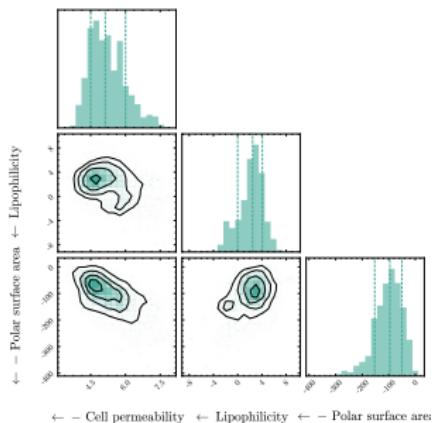


⁷Joe, *Multivariate Models and Dependence Concepts* (1997).

Model-based Pareto front

Domain knowledge or information from unpaired observations of Y (without X associations) can be encoded in the choices of

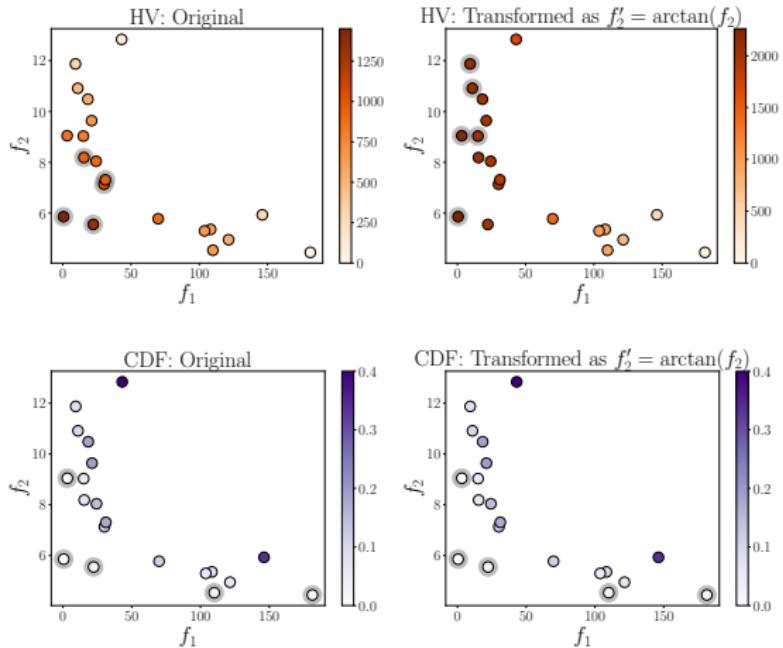
- ▶ marginal distributions
- ▶ pair copula models
- ▶ vine structure



$$f(x_1, x_2, x_3, x_4) = c_{14|23} \cdot c_{13|2} \cdot c_{24|3} \cdot c_{12} \cdot c_{23} \cdot c_{34} \cdot f_1(x_1) \cdot f_2(x_2) \cdot f_3(x_3) \cdot f_4(x_4).$$

Desirable invariance properties

CDF is invariant to arbitrary monotonic transformations of objectives, while HV is very sensitive. Important for common unit conversions (e.g., linear $\mu\text{m} \rightarrow \text{nm}$, loglike KD \rightarrow pKD to remove tails)!



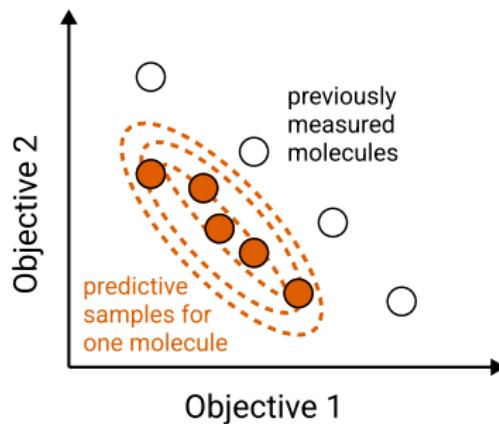
Quality indicators to MOBO acquisition functions

Quality indicator $I : 2^{\mathcal{Y}} \rightarrow \mathbb{R}$ scores **already-measured** sets of molecules.

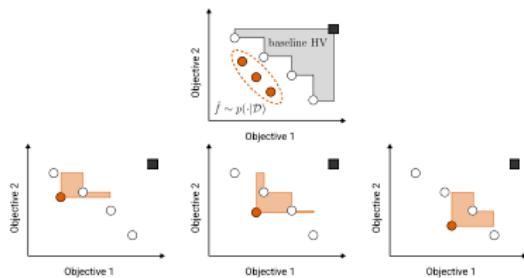
→ How well did we exploit?

Acquisition function $a^{\hat{f}} : \mathcal{X} \rightarrow \mathbb{R}$ scores each molecule based on **predictions** by the surrogate \hat{f} .

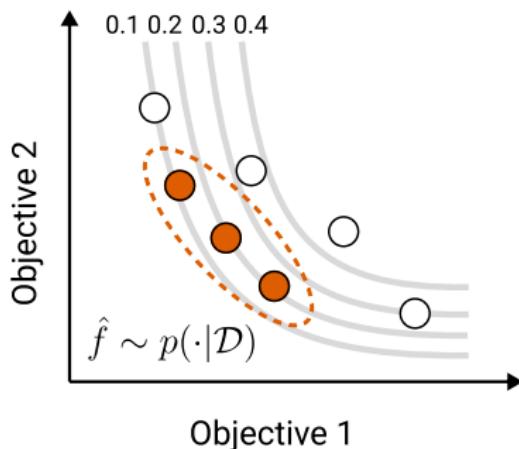
→ How can we balance exploration with exploitation?



Quality indicators to MOBO acquisition functions



HV indicator \rightarrow expected hypervolume improvement (EHVI)



CDF indicator \rightarrow BOtied

Emmerich, Deutz, and Klinkenberg,
"Hypervolume-based expected improvement:
Monotonicity properties and exact
computation" (2011)

Content

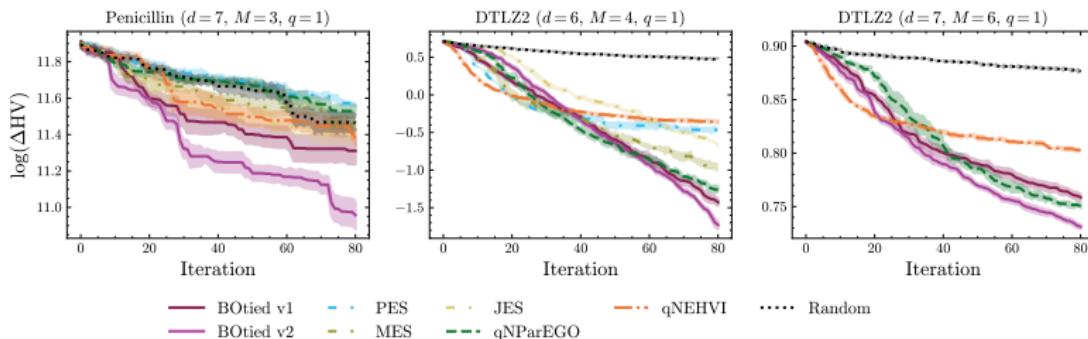
Background Method

- ▶ Connection between the CDF ranks and the Pareto front
- ▶ BOtied: MOBO based on the CDF

Empirical results

Empirical results

BOtied outperforms EHVIs on standard synthetic benchmark problems for MOBO, even in terms of HV.

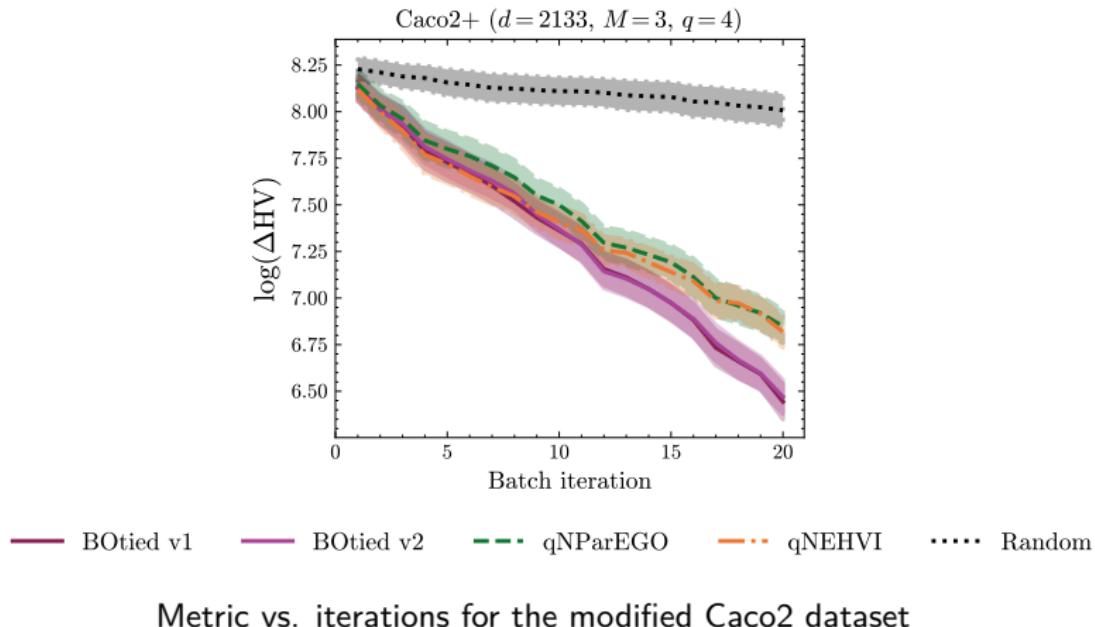


Metric vs. iterations for two synthetic problems.

Metric: $\log(\Delta\text{HV}) := \log(HV(\mathcal{P}) - HV(\hat{\mathcal{P}}))$ (lower is better)

Empirical results

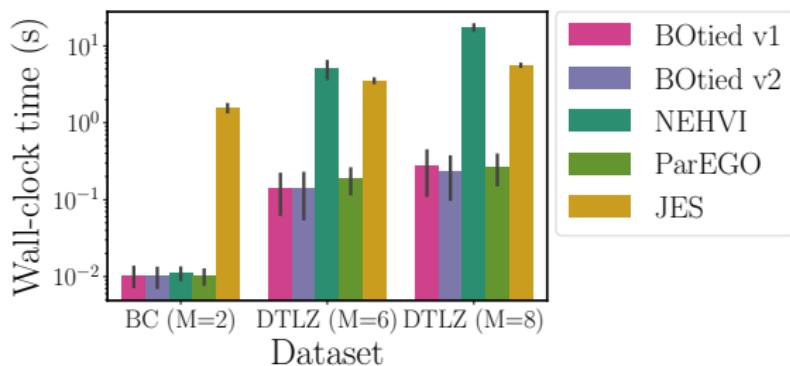
BOtied outperforms EHVI on a real-world dataset of cell permeability measurements.



Computational efficiency

- ▶ Vine copula implementation makes BOtied very fast relative to EHVI and joint entropy search (JES), both involving M -dim integrals
- ▶ BOtied has competitive wall-clock time with ParEGO, which randomly scalarizes the objectives (effectively $M = 1$)

Per function evaluation:



Summary: BOtied

BOtied is an acquisition function well suited for the joint optimization of multiple biophysical properties in active molecular design.

- ▶ efficiently implemented using vine copulas for $M > 4$ properties
- ▶ invariant to monotonic transformations of property values
- ▶ enables integration of domain knowledge in model-based construction of Pareto front

Framework is general: hierarchical Bayesian inference, mixed-variable outcomes, **differentiable BOtied**, integration into generative models for guided generation

- ▶ **Park, J.W.***, Tagasovska, N.*, Maser, M., Ra, S., and Cho, K. “BOtied: Multi-objective Bayesian optimization with tied multivariate ranks.” ICML (2024). arXiv: 2306.00344
- ▶ **Park, J.W.**, Tibshirani, R., and Cho, K. “Semiparametric conformal prediction.” AISTATS (2025). arXiv: [2411.02114](https://arxiv.org/abs/2411.02114)

Motivation

- ▶ Many applications require prediction sets spanning multiple correlated targets.
 - ▶ Example: small molecule ADME characterization involves ~ 50 endpoints with similar assays repeated across species. We require uncertainties for lab prioritization.
- ▶ Consider a multi-target regression task given a dataset $\{(X^{(i)}, Y^{(i)})\}_{i \in \mathcal{I}}$ of input features $X^{(i)} \in \mathcal{X}$ and labels $Y^{(i)} \in \mathcal{Y}$, viewed as $|\mathcal{I}|$ exchangeable samples drawn from $P_{XY} = P_x \times P_{Y|X}$.
- ▶ Given a miscoverage level α , **conformal prediction (CP)** produces *marginally valid* prediction sets $\Gamma_{1-\alpha}$ with minimal assumptions.

Marginal validity⁸

A set $\Gamma_{1-\alpha}(X^*)$ is *marginally valid* if it contains the true response Y^* w.p. at least $1 - \alpha$:

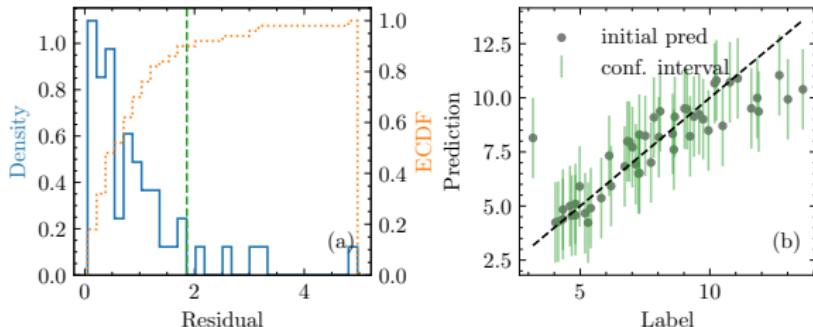
$$\mathbb{P}[Y^* \in \Gamma_{1-\alpha}(X^*)] \geq 1 - \alpha.$$

⁸Weaker condition than *conditional validity*, $\mathbb{P}[Y^* \in \Gamma_{1-\alpha}(X^*) | X^*] \geq 1 - \alpha$

Split conformal prediction⁹

1. Split data into proper training data $\mathcal{I}_{\text{train}}$ and calibration data \mathcal{I}_{cal} .
2. Fit the *underlying predictor* $\hat{f} : \mathcal{X} \rightarrow \mathcal{Y}$ on $\mathcal{I}_{\text{train}}$.
3. Define the *non-conformity score* (e.g., $V(X, Y, \hat{f}) = |Y - \hat{f}(X)|$) and evaluate it on \mathcal{I}_{cal} of size n .
4. Given target level α , with $Q_{1-\alpha}$ defined as the $\lceil(1 - \alpha)(n + 1)\rceil$ -th smallest of the scores, return the conformalized prediction set:

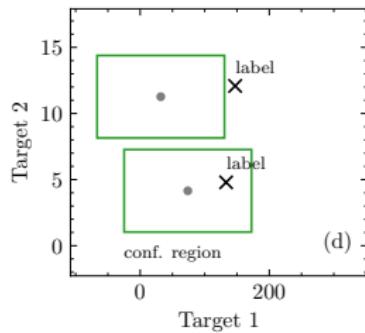
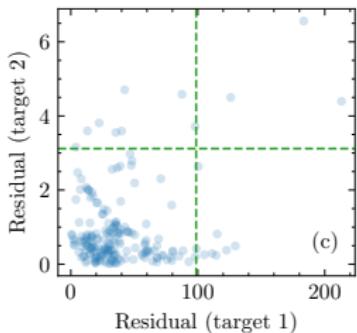
$$\Gamma_{1-\alpha}(X^*) = \{Y : V(X^*, Y, \hat{f}) \leq Q_{1-\alpha}\}.$$



⁹I'll present the method in terms of split (inductive) CP for simplicity, but it applies to full (transductive) and CV, jackknife variants too.

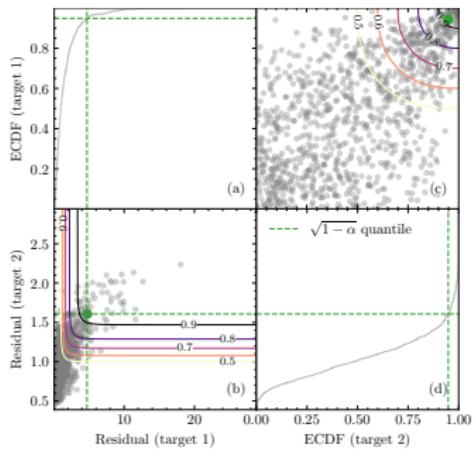
Naive extension to multiple targets

- ▶ Why not apply CP independently to each target?
- ▶ Marginal level $1 - \alpha_j$ for each target j becomes more stringent in order to satisfy the global coverage level of $1 - \alpha$.
 - ▶ e.g., $1 - \alpha_j = \sqrt{1 - \alpha}$ for $d = 2$
 - ▶ $(1 - \alpha)^{1/d} \rightarrow 1$: leads to large prediction sets as d increases
 - ▶ n may not be large enough to accommodate extreme $1 - \alpha_j$ levels
- ▶ Joint modeling can yield more efficient (tighter) prediction sets.



Multivariate quantiles for score vectors

- Canonical ordering does not exist in \mathbb{R}^d for $d > 1$.¹⁰ → Estimate the joint cumulative distribution function (CDF) of the scores, $F(s) = \mathbb{P}[S_1 \leq s_1, \dots, S_d \leq s_d] = \mathbb{P}[S \preccurlyeq s]$, where $s \in \mathbb{R}^d$ using nonparametric vine copulas.
- Obtain the quantile as its generalized inverse, $F^{-1}(p) = \{s \in \mathbb{R}^d : F(s) = p\}$.



¹⁰Koltchinskii, “M-estimation, convexity and quantiles” (1997).

Semiparametric one-step correction

We perform flexible density estimation to obtain \hat{F} . But the CP algorithm only requires its low-dimensional functional, the $1 - \alpha$ quantile $Q_{1-\alpha}$.

- ▶ When estimating a functional $\Psi(F)$ of the unknown distribution F ...
- ▶ a plug-in estimator $\Psi(\hat{F})$ is often biased.¹¹
 - ▶ c.f. “appeal” of Bayes to model all nuisance variables
- ▶ We can debias the plug-in using the **efficient influence function**, which captures the sensitivity of $\Psi(F)$ to changes in F .

¹¹Tsiatis, *Semiparametric theory and missing data* (2006).

Semiparametric CP algorithm

Algorithm Semiparametric Conformal Prediction

- 1: **Input:** Labeled data \mathcal{I} , test inputs $\mathcal{I}_{\text{test}}$, target coverage level $1 - \alpha$
- 2: **Output:** Prediction set $\Gamma_{1-\alpha}(X^*)$ for test input X^*
- 3: Split \mathcal{I} into $\mathcal{I}_{\text{train}}$ and \mathcal{I}_{cal}
- 4: Train the underlying algorithm \hat{f} on $\mathcal{I}_{\text{train}}$
- 5: Evaluate vector scores: $S_j^{(i)} \leftarrow V(X_j^{(i)}, Y_j^{(i)}, \hat{f}) \quad \forall i \in \mathcal{I}_{\text{cal}}, j \in [d]$
- 6: **Estimate the score distribution using the vine copula:**
 1. Compute the marginal ECDF $\hat{F}_j \quad \forall j \in [d]$
 2. Get uniform marginals $U_j^{(i)} \leftarrow \hat{F}_j(S_j^{(i)})$
 3. Fit the copula \hat{C} on $U^{(i)}$
- 7: **Optimize for quantile:** $U^* \leftarrow \arg \min_{U \in [0,1]^d} \|U\|_1 \text{ s.t. } \hat{C}(U) \geq 1 - \alpha$
- 8: **One-step correction:** $U_{1\text{-step}} \leftarrow U^* + \frac{1}{n} \sum_{i=1}^n \psi_{\hat{C}}(U^{(i)})$
- 9: Mapping back to score space: $Q_{1-\alpha} \leftarrow [\hat{F}_1^{-1}(U_1^*), \dots, \hat{F}_d^{-1}(U_d^*)]$
- 10: **Return:** $\Gamma_{1-\alpha}(X^*) = \{Y \in \mathbb{R}^d : V(X^*, Y, \hat{f}) \leq Q_{1-\alpha}\}$, where $v \leq w$ for $v, w \in \mathbb{R}^d$ if $v_1 \leq w_1, \dots, v_d \leq w_d$.

Theoretical Guarantees

Theorem (Asymptotic exact coverage)

Our prediction set $\Gamma_{1-\alpha}(X^)$ satisfies, for a test point X^*, Y^* ,*

$$\mathbb{P}[Y^* \in \Gamma_{1-\alpha}(X^*)] \rightarrow 1 - \alpha \quad \text{as } n \rightarrow \infty.$$

→ Proof follows from consistency of the copula estimator (and thus its quantile) and asymptotic normality of the one-step estimator.

Theorem (Approximate validity)

Suppose the total variation distance between F and \hat{F} is bounded by ϵ . That is, $\sup_S |F(S) - \hat{F}(S)| \leq \epsilon$. Then our prediction set $\Gamma_{1-\alpha}$ is marginally valid at the $1 - \alpha - \epsilon$ level:

$$\mathbb{P}[Y^* \in \Gamma_{1-\alpha}(X^*)] \geq 1 - \alpha - \epsilon,$$

with or without the one-step correction.

→ TV distance between F^* and \hat{F} upper-bounds the TV distance between their quantiles as well as that between the quantile of F and the one-step-corrected quantile of \hat{F} .

Experimental Setup

- ▶ **Task:** Multi-target regression
- ▶ **Datasets:** Synthetic ($d = 3, n = 96$) and several real-world datasets with $d \in \{6, 8, 16\}$.
- ▶ **Underlying predictor:** Multi-task Lasso point predictor¹² (or conditional density estimator, in the Appendix).
- ▶ **Metrics:**
 - ▶ *Coverage:* Empirical frequency of the true label in the prediction set.
 - ▶ *Efficiency:* (Log-)Volume of the prediction set (smaller is better).

¹²Tibshirani, “Regression shrinkage and selection via the lasso” (1996). A set of small, light-blue navigation icons typically used in Beamer presentations for navigating between slides and sections.

Comparison

- ▶ **Independent:** univariate calibration applied independently to each target at the $(1 - \alpha)^{1/d}$ level
- ▶ **Scalar score:** calibration applied to a scalar score defined as the L_2 norm of the prediction error $V(X, Y, \hat{f}) = \|Y - \hat{f}(X)\|_2^{13}$
- ▶ **Empirical copula:** fit on vector scores with the constraint that $U_1^* = \dots = U_d^*$ ¹⁴
- ▶ **Proposed Methods (Plug-in and Corrected):** Yield nearly exact coverage and improved efficiency.

¹³Yields prediction sets shaped as d -dimensional balls. The L_1 norm would yield cross-polytopes (d -dimensional generalization of diamonds)

¹⁴Messoudi, Destercke, and Rousseau, “Copula-based conformal prediction for multi-target regression” (2021).

One-step correction helps

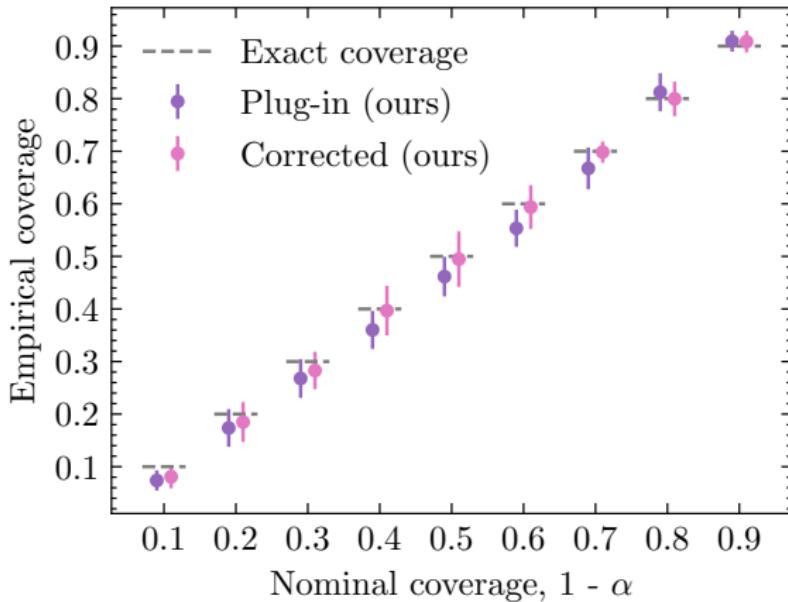


Figure: Penicillin production simulator dataset¹⁵ with $d = 3, n = 96$

¹⁵Liang and Lai, "Scalable bayesian optimization accelerates process optimization of penicillin production" (2021).

Empirical copula has high variance

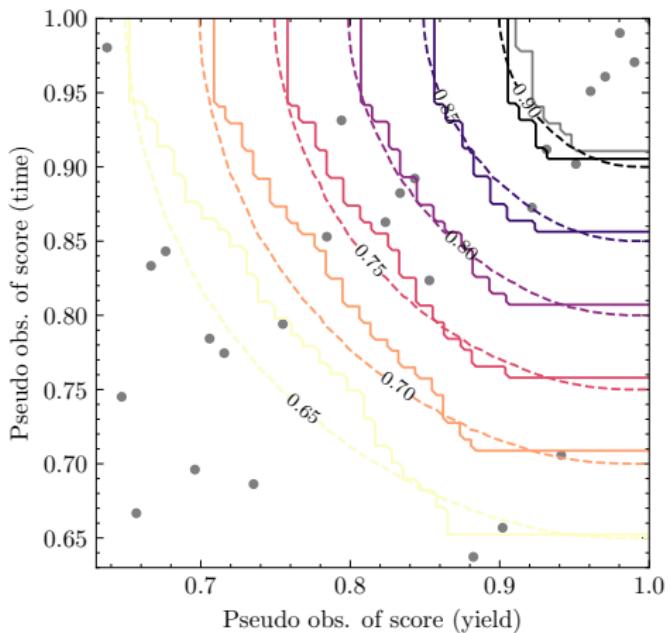


Figure: At the 0.9 level, the estimated curve (dashed black) falls between the true curve (solid black) and empirical (solid gray) curve computed from 96 points, some shown in gray dots.

Results on real-world datasets

Table: Mean \pm standard error across five seeds. Target coverage is 0.9.

Method	Stock ¹⁶ ($d = 6, n = 63$)		Caco2 ¹⁷ ($d = 6, n = 137$)		rf1 ¹⁸ ($d = 8, n = 225$)	
	Coverage	Efficiency ↓	Coverage	Efficiency ↓	Coverage	Efficiency ↓
Independent	0.90 \pm 0.01	-2.4 \pm 0.2	0.95 \pm 0.01	12.2 \pm 0.1	0.96 \pm 0.01	29.2 \pm 0.5
Scalar score	0.90 \pm 0.01	-1.8 \pm 0.3	0.92 \pm 0.01	28.3 \pm 0.1	0.92 \pm 0.00	27.7 \pm 0.3
Empirical copula	0.50 \pm 0.05	-4.7 \pm 0.5	0.42 \pm 0.08	8.4 \pm 0.4	0.42 \pm 0.12	21.2 \pm 2.6
Plug-in (ours)	0.87 \pm 0.02	-2.9 \pm 0.2	0.90 \pm 0.01	11.0 \pm 0.1	0.91 \pm 0.01	25.0 \pm 0.2
Corrected (ours)	0.90 \pm 0.02	-2.8 \pm 0.2	0.93 \pm 0.01	11.5 \pm 0.2	0.91 \pm 0.01	25.1 \pm 0.3
Method	rf2 ($d = 8, n = 225$)		scm1d ($d = 16, n = 448$)		scm20d ($d = 16, n = 448$)	
	Coverage	Efficiency ↓	Coverage	Efficiency ↓	Coverage	Efficiency ↓
Independent	0.96 \pm 0.01	29.2 \pm 0.5	0.96 \pm 0.01	114.1 \pm 0.4	0.96 \pm 0.01	114.1 \pm 0.4
Scalar score	0.92 \pm 0.01	27.7 \pm 0.3	0.89 \pm 0.01	109.0 \pm 0.3	0.89 \pm 0.01	109.0 \pm 0.3
Empirical copula	0.42 \pm 0.12	21.2 \pm 2.6	0.75 \pm 0.05	108.5 \pm 0.8	0.75 \pm 0.05	108.5 \pm 0.8
Plug-in (ours)	0.90 \pm 0.01	25.0 \pm 0.2	0.92 \pm 0.01	111.4 \pm 0.1	0.92 \pm 0.01	111.4 \pm 0.2
Corrected (ours)	0.91 \pm 0.01	25.1 \pm 0.3	0.91 \pm 0.01	110.7 \pm 0.2	0.90 \pm 0.01	110.6 \pm 0.2

¹⁶Liu and Yeh, "Using mixture design and neural networks to build stock selection decision support systems" (2017).

¹⁷Wang et al., "ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting" (2016); Park et al., "BOtied: Multi-objective Bayesian optimization with tied multivariate ranks" (2023).

¹⁸Spyromitros-Xioufis et al., "Multi-target regression via input space expansion: treating targets as inputs" (2016).

Unlocking missing-at-random (MAR) data

Training and calibration data often have missing observations, especially as d gets larger. If we only take instances for which we observe both, we end up with a **biased** quantile estimate. We can do missingness imputation with copulas¹⁹.

Common scenario: target labels are missing at random (MAR), such that the labels for target 2 are only observed when target 1 observations exceed a certain value.

¹⁹Feldman and Kowal, "Nonparametric Copula Models for Multivariate, Mixed, and Missing Data" (2024).

Summary: semiparametric conformal prediction

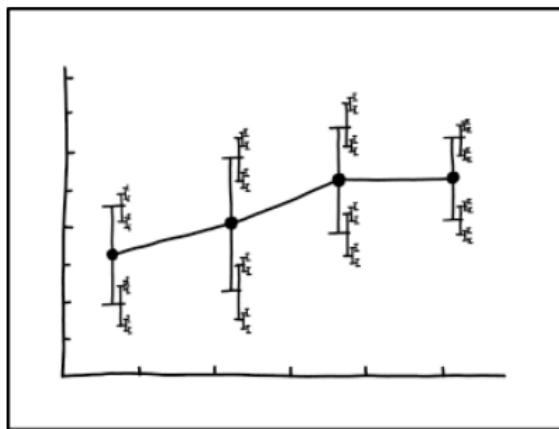
- ▶ We introduced the semiparametric conformal calibration scheme, adapted for design settings with **many correlated molecular properties**.
- ▶ By combining nonparametric vine copulas with a one-step estimator, our method yields **efficient** prediction sets by modeling the tails near the $1 - \alpha$ **joint quantile** of interest.
 - ▶ Baselines tend to overcover or suffer from high variance in the tails.
- ▶ It guarantees **asymptotically exact** coverage and approximate validity in finite samples.
- ▶ A particular copula model allows working with **missing-at-random** observations.

Power of modeling the tails in AI-driven optimization

Surrogate function (multivariate regression) → $p_{Y|X}$

► BOtied population distribution (density estimation) → p_Y

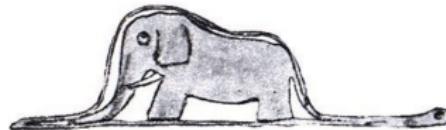
► SemiCP population distribution (density estimation) → $p_{V(X,Y,\hat{f})}$



I DON'T KNOW HOW TO PROPAGATE
ERROR CORRECTLY, SO I JUST PUT
ERROR BARS ON ALL MY ERROR BARS.

Thank you!

🌐 jiwoncpark.github.io 🐦 [jiwoncpark](https://twitter.com/jiwoncpark) 📩 park.ji_won@gene.com



References |

-  Jain, Tushar et al. "Biophysical properties of the clinical-stage antibody landscape". In: *Proceedings of the National Academy of Sciences* 114.5 (2017), pp. 944–949.
-  Wang, Ning-Ning et al. "ADME properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting". In: *Journal of Chemical Information and Modeling* 56.4 (2016), pp. 763–773.
-  Konakovic Lukovic, Mina, Yunsheng Tian, and Wojciech Matusik. "Diversity-guided multi-objective bayesian optimization with batch evaluations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 17708–17720.
-  Emmerich, Michael TM, André H Deutz, and Jan Willem Klinkenberg. "Hypervolume-based expected improvement: Monotonicity properties and exact computation". In: *2011 IEEE Congress of Evolutionary Computation (CEC)*. IEEE. 2011, pp. 2147–2154.
-  Yang, Kaifeng et al. "A multi-point mechanism of expected hypervolume improvement for parallel multi-objective bayesian global optimization". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2019, pp. 656–663.
-  Binois, Mickaël, Didier Rullière, and Olivier Roustant. "On the estimation of Pareto fronts from the point of view of copula theory". In: *Information Sciences* 324 (2015), pp. 270–285.
-  Joe, Harry. *Multivariate Models and Dependence Concepts*. Chapman & Hall/CRC, 1997.

References II

-  Koltchinskii, Vladimir I. "M-estimation, convexity and quantiles". In: *The annals of Statistics* (1997), pp. 435–477.
-  Tsiatis, Anastasios A. *Semiparametric theory and missing data*. Vol. 4. Springer, 2006.
-  Tibshirani, Robert. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1 (1996), pp. 267–288.
-  Messoudi, Soundouss, Sébastien Destercke, and Sylvain Rousseau. "Copula-based conformal prediction for multi-target regression". In: *Pattern Recognition* 120 (2021), p. 108101.
-  Liang, Qiaohao and Lipeng Lai. "Scalable bayesian optimization accelerates process optimization of penicillin production". In: *NeurIPS 2021 AI for Science Workshop*. 2021.
-  Liu, Yi-Cheng and I-Cheng Yeh. "Using mixture design and neural networks to build stock selection decision support systems". In: *Neural Computing and Applications* 28 (2017), pp. 521–535.
-  Park, Ji Won et al. "BOtied: Multi-objective Bayesian optimization with tied multivariate ranks". In: *arXiv preprint arXiv:2306.00344* (2023).
-  Spyromitros-Xioufis, Eleftherios et al. "Multi-target regression via input space expansion: treating targets as inputs". In: *Machine Learning* 104 (2016), pp. 55–98.
-  Feldman, Joseph and Daniel R Kowal. "Nonparametric Copula Models for Multivariate, Mixed, and Missing Data". In: *Journal of Machine Learning Research* 25.164 (2024), pp. 1–50.

References III

-  Hansen, Nikolaus. "The CMA evolution strategy: a comparing review". In: *Towards a new evolutionary computation: Advances in the estimation of distribution algorithms* (2006), pp. 75–102.
-  Park, Ji Won et al. "PropertyDAG: Multi-objective Bayesian optimization of partially ordered, mixed-variable properties for biological sequence design". In: *arXiv preprint arXiv:2210.04096* (2022).
-  Tagasovska, Natasa and Park, Ji Won et al. "Antibody DomainBed: Out-of-Distribution Generalization in Therapeutic Protein Design". In: (2023).
-  Huard, David, Guillaume Evin, and Anne-Catherine Favre. "Bayesian copula selection". In: *Computational Statistics & Data Analysis* 51.2 (2006), pp. 809–822.
-  Belakaria, Syrine, Aryan Deshwal, and Janardhan Rao Doppa. "Max-value entropy search for multi-objective Bayesian optimization". In: *Advances in neural information processing systems* 32 (2019).
-  Ichimura, Hidehiko and Whitney K Newey. "The influence function of semiparametric estimators". In: *Quantitative Economics* 13.1 (2022), pp. 29–61.
-  Hines, Oliver et al. "Demystifying statistical learning based on efficient influence functions". In: *The American Statistician* 76.3 (2022), pp. 292–304.
-  Koenker, Roger and Gilbert Bassett Jr. "Regression quantiles". In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.

References IV

-  Yiu, Andrew et al. "Semiparametric posterior corrections". In: *arXiv preprint arXiv:2306.06059* (2023).
-  Le Cam, Lucien. "On the asymptotic theory of estimation and testing hypotheses". In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 3. University of California Press. 1956, pp. 129–157.
-  Pfanzagl, J and J Pfanzagl. "Existence of Asymptotically Efficient Estimators for Functionals". In: *Contributions to a General Asymptotic Statistical Theory* (1982), pp. 196–210.
-  Newey, Whitney K, Fushing Hsieh, and James Robins. "Undersmoothing and bias corrected functional estimation". In: *Working Paper* (1998).