PropertyDAG: Multi-Objective Bayesian Optimization for Biological Sequence Design

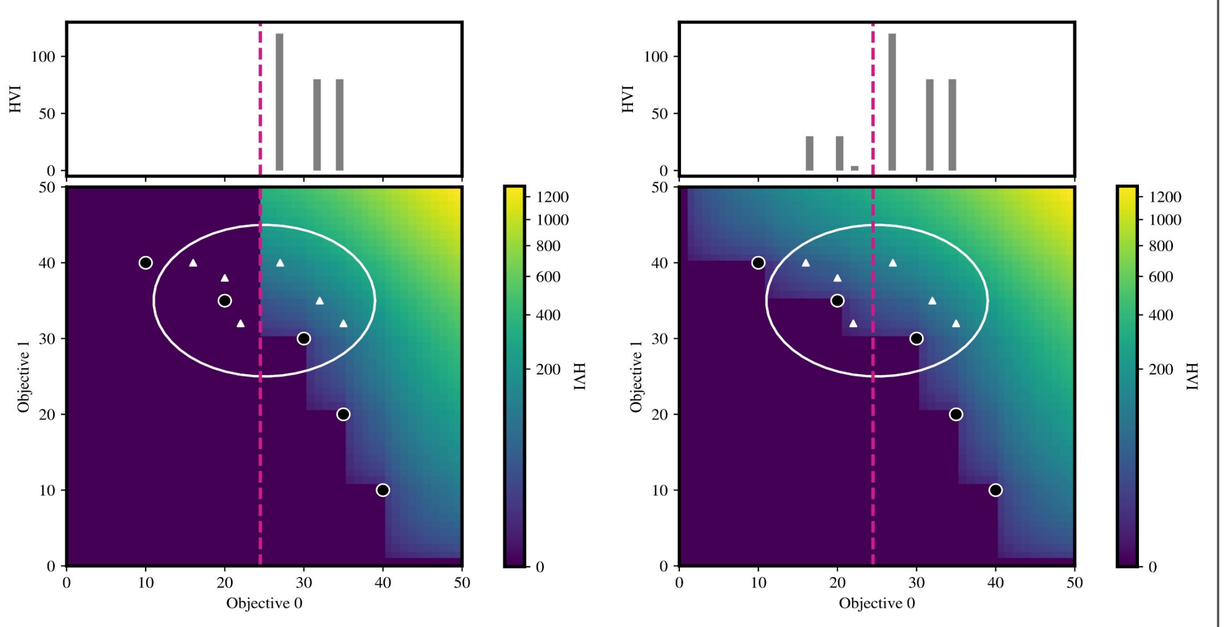
Ji Won Park (park.ji_won@gene.com), Samuel Stanton, Saeed Saremi, Andy Watkins, Henri Dwyer, Vladimir Gligorijević, Rich Bonneau, Stephen Ra, Kyunghyun Cho

Prescient Design
A Genentech Accelerator

Motivation

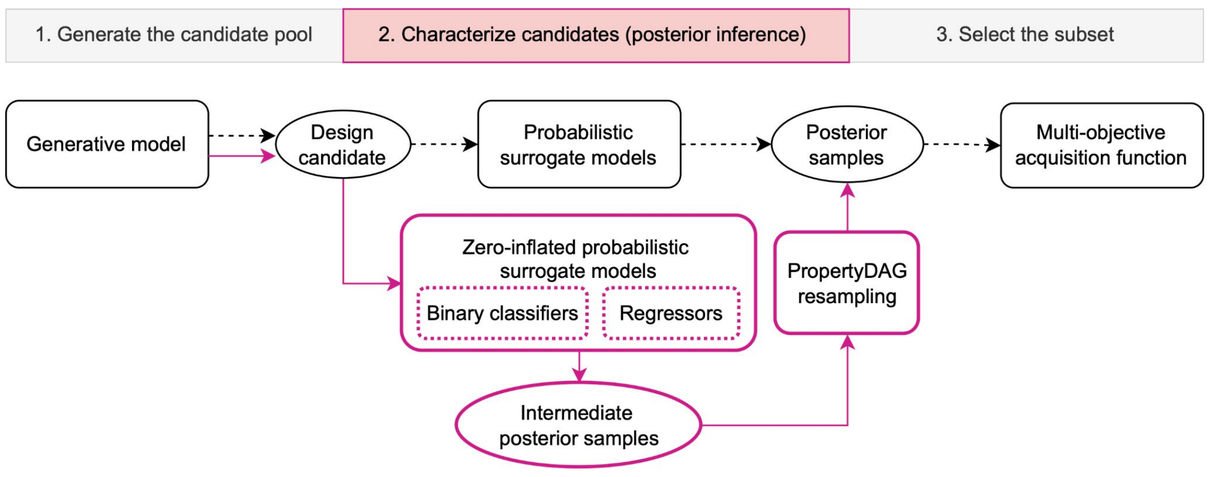
- *In silico* drug design involves fulfilling property desiderata such as expression/synthesizability, potency against a therapeutic target, and various developability properties [1, 2].
- Sequential nature of wet-lab characterization inspires optimization conditioned on preceding properties satisfying some constraints.
- Multi-objective Bayesian optimization (BO) offers a principled framework for navigating the exploration-exploitation trade-off in design space across multiple properties.
- Acquisition functions like expected hypervolume improvement (EHVI) [3] provide Pareto fronts with excellent coverage [4] but do not allow discrimination between different regions on the Pareto frontier.

PropertyDAG-EHVI



Say we want designs that maximize Objective 1 and exceed a threshold (dashed **magenta**) in Objective 0, given some baseline (**black** dots). Consider six samples (white triangles) from the posterior (white contour). PropertyDAG transforms the posterior samples below the threshold so that their HVI contribution is zero.

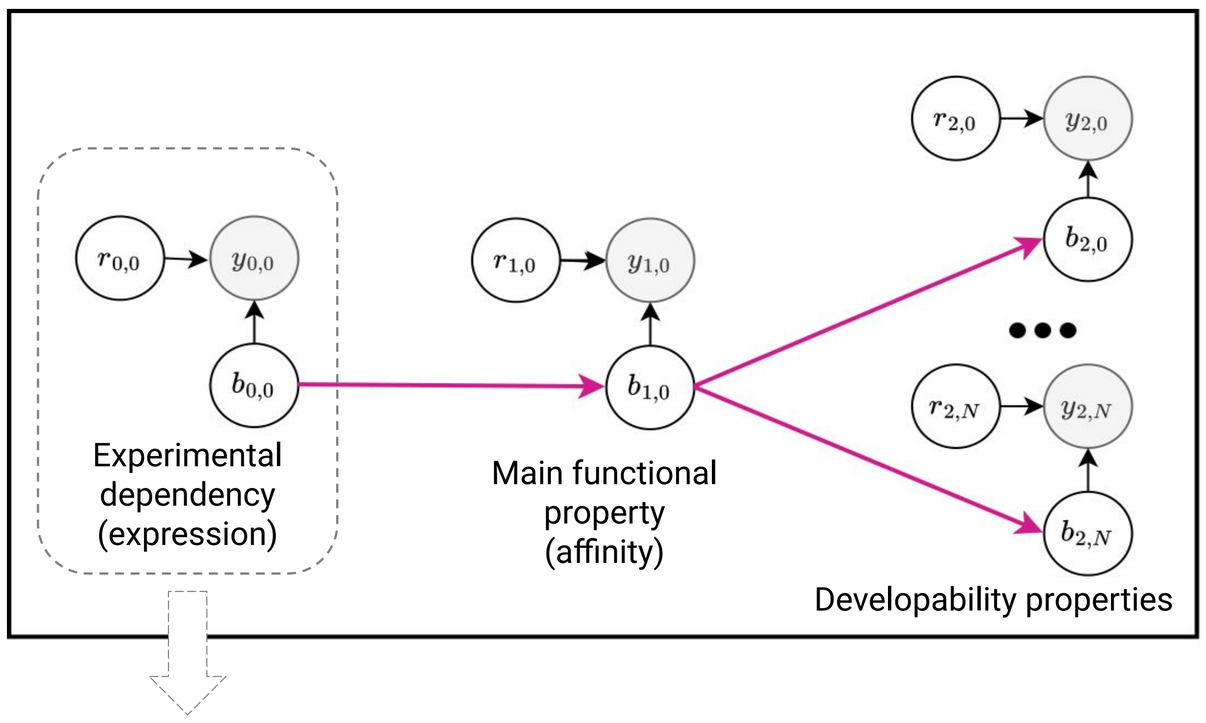
Method



PropertyDAG-BO pipeline. Traditional multi-objective BO pipeline (dashed **black**) compared to PropertyDAG-BO (**magenta**) modifying the posterior inference step.

1. Choosing a PropertyDAG

Assign *experimental dependencies* (*e.g.*, expression → affinity) or *prioritized properties* (*e.g.*, affinity → specificity, stability) as parents.



2. Zero-inflated modeling with PropertyDAG

Characterize each property with a probabilistic *classifier* (*e.g.*, does the antibody bind to the antigen?) and a probabilistic *regressor for the positives only* (*e.g.*, what is the affinity, if it does bind?).

Modify their posterior distributions so that a given property is zero if any of its predecessor properties is zero:

$$p(b_k \mid \boldsymbol{x}, \mathcal{D}_t) = \begin{cases} 0 & \text{if } \exists j \in \text{pred}(k) \text{ s.t. } b_j = 0, \\ p(b_k \mid \boldsymbol{x}, \mathcal{D}_t, \theta_b) & \text{else,} \end{cases}$$

Classifier

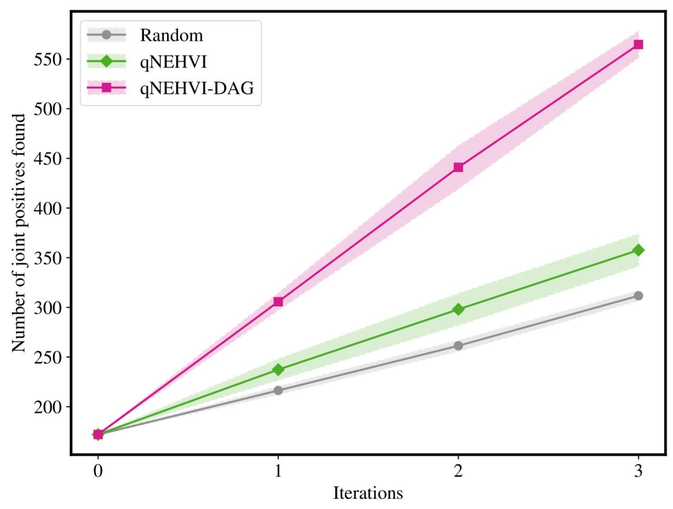
$$p(\hat{f}_k(\boldsymbol{x}) = c \mid \mathcal{D}_t) = \begin{cases} p(b_k = 0 \mid \boldsymbol{x}, \mathcal{D}_t) & \text{if } c = 0, \\ p(b_k = 1 \mid \boldsymbol{x}, \mathcal{D}_t)\, p(r_k = c \mid \boldsymbol{x}, \mathcal{D}_t, \theta_r) & \text{else,} \end{cases}$$

Regressor

Experiments

PropertyDAG for antibody design: expression (Obj 0, binary) → affinity (Obj 1, zero-inflated, continuous).
We simulate 3 iterations of active learning by splitting a dataset of antibody scFv sequences and associated expression and affinity labels into 5 groups.

Initial training set (N = 1,230)	**Pool 1** (N = 736)	**Pool 2** (N = 746)	**Pool 3** (N = 711)	**Test set** (N = 600)
Initializes surrogates	Select 200 in iteration 1 from this pool	Select 200 in iteration 2 from this pool	Select 200 in iteration 3 from this pool	Held out, for final evaluation

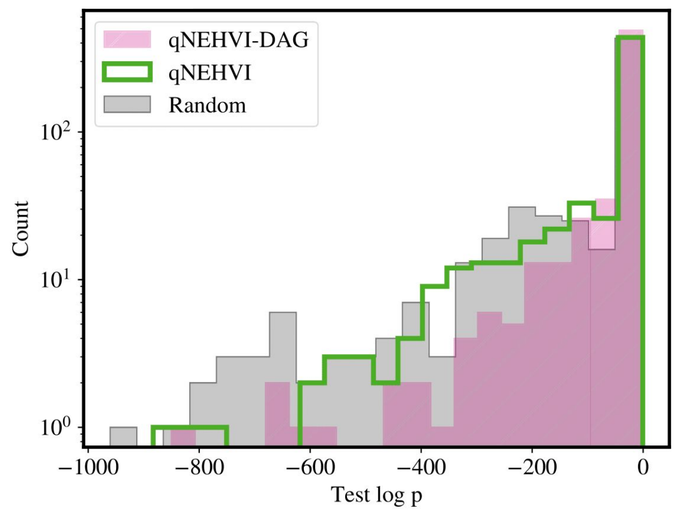


Joint positives identified over iterations

PropertyDAG-EHVI selection (**magenta**) identifies significantly more expressing binders relative to standard EHVI (**green**) and random (**gray**) selections.

Log posterior density of joint positives in test set

Surrogate models from PropertyDAG-EHVI (**magenta**) selection has the most accurate beliefs about the joint positives after the final iteration, relative to standard EHVI (**green**) and random (**gray**) selections.



Summary

- PropertyDAG sits on top of multi-objective BO to make it amenable to a common scenario in drug design, where a hierarchical structure, or partial ordering, exists among the objectives.
- PropertyDAG-BO can identify significantly more designs that are jointly positive (*i.e.*, exceeding a chosen threshold in all properties) than can standard BO.

References

[1] Jain *et al.* (2017). Biophysical properties of the clinical antibody landscape. *PNAS,* 114(5), 944-949

[2] Raybould *et al*. (2019). Five computational developability guidelines for therapeutic antibody profiling. *PNAS,* 116(10) 4025-4030

[3] Emmerich (2005). Single- and multi-objective evolutionary design optimization assisted by Gaussian random field metamodels. Doctoral dissertation, Dortmund, Univ., Diss., 2005

[4] Zitzler *et al.* 2004. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on evolutionary computation*, 7(2), 117-132.