Prescient Design

A Genentech Accelerator

PROPerty ENhancer: Match your data to follow the gradient

Nataša Tagasovska, Vlad Gligorijevic, Kyunghyun Cho, Andreas Loukas




Genentech

A Member of the Roche Group

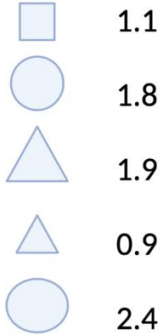
Design Optimization in Low-Data Regimes

- Problem Setup and Motivation

Prescient Design
A Genentech Accelerator

Dataset

(x, y)



Task: Increase the value of y

Applications

x	y
polygon	area
molecule	functional property
portfolio	revenue
airfoil	aerodynamics

Genentech
A Member of the Roche Group

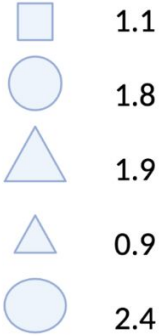
Design Optimization in Low-Data Regimes

- Problem Setup and Motivation

Prescient Design
A Genentech Accelerator

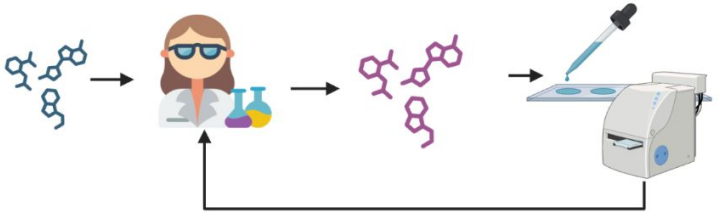
Dataset

(x, y)



Task: Increase the value of y

Execution: Domain expert design cycles



Applications

x	y
polygon	area
molecule	functional property
portfolio	revenue
airfoil	aerodynamics

Genentech
A Member of the Roche Group

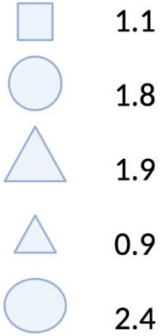
Design Optimization in Low-Data Regimes

- **Problem Setup and Motivation**

Prescient Design
A Genentech Accelerator

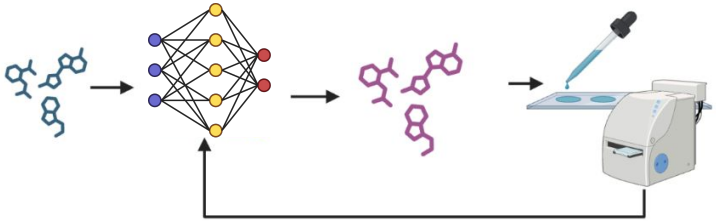
Dataset

(x, y)



Task: Increase the value of y

Execution: Domain expert design cycles



Applications

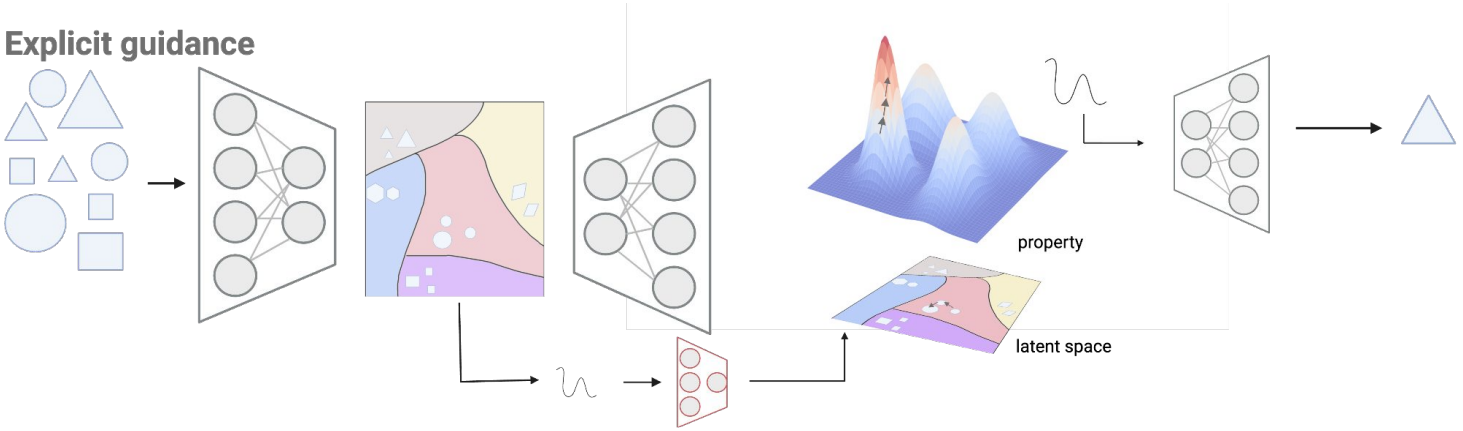
x	y
polygon	area
molecule	functional property
portfolio	revenue
airfoil	aerodynamics

Genentech
A Member of the Roche Group

Design Optimization in Low-Data Settings

Prescient Design
A Genentech Accelerator

- **Challenges in explicitly guided design: generative + discriminative model**

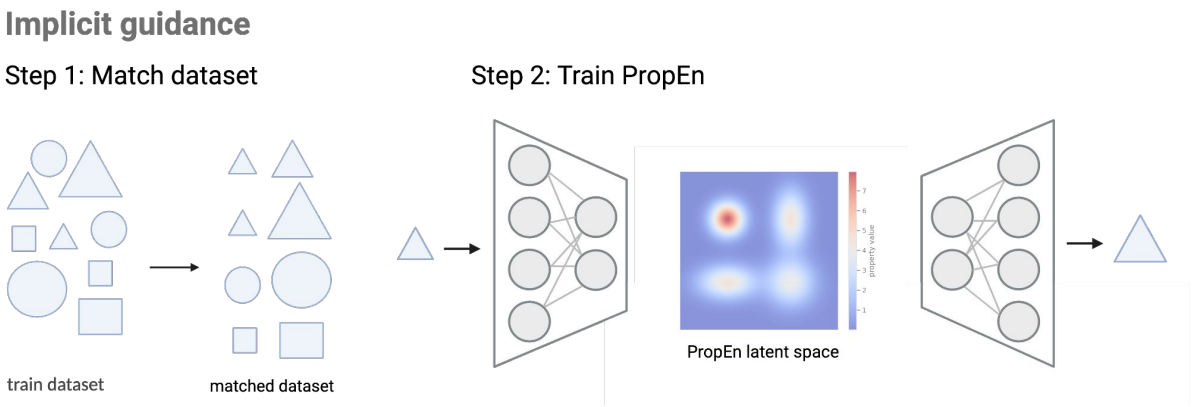


- Requires a (trustworthy) discriminative model
- Large training datasets
- Falls off data-manifold
- Difficulty in non-convex, complex distributions

Design Optimization in Low-Data Settings

- **Implicitly guided design**

Prescient Design
A Genentech Accelerator

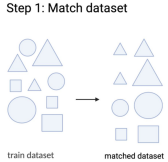


- No need for discriminative model
- Low data regime
- In-distribution designs (with theoretical guarantees)
- Linear approximation of the gradient close to starting designs

Genentech
A Member of the Roche Group

Property Enhancer - PropEn

- Step 1: Match your dataset



Prescient Design
A Genentech Accelerator

We view the group of samples with superior property values as the **treated** group and their lower value counterpart as the **control** group. This motivates us to construct a **matched dataset** for every *(x, y)* within *D*:

$$\mathcal{M} = \left\{ (x, x') \;\middle|\; \begin{array}{l} x, x' \in \mathcal{D} \\ \|x' - x\|^2 \leq \Delta_x,\; g(x') - g(x) \in (0, \Delta_y] \end{array} \right\},$$

Where Δ_x and Δ_y are predefined, positive thresholds that will trade-off exploration vs exploitation.

One control - to - many treatments -> extending dataset by large order of magnitude!

Genentech
A Member of the Roche Group

Property Enhancer - PropEn

- Step 2: Approximate the gradient

Prescient Design
A Genentech Accelerator

Once a dataset has been matched, we train a deep encoder-decoder network f_θ over M by minimizing the **matched reconstruction objective**:

$$\ell(f_\theta; \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{(x,x') \in \mathcal{M}} \ell(f_\theta(x), x')$$

Minimizing the matched reconstruction objective yields a model that **approximates the direction of the gradient** of g(·), even if no property predictor has been explicitly trained.

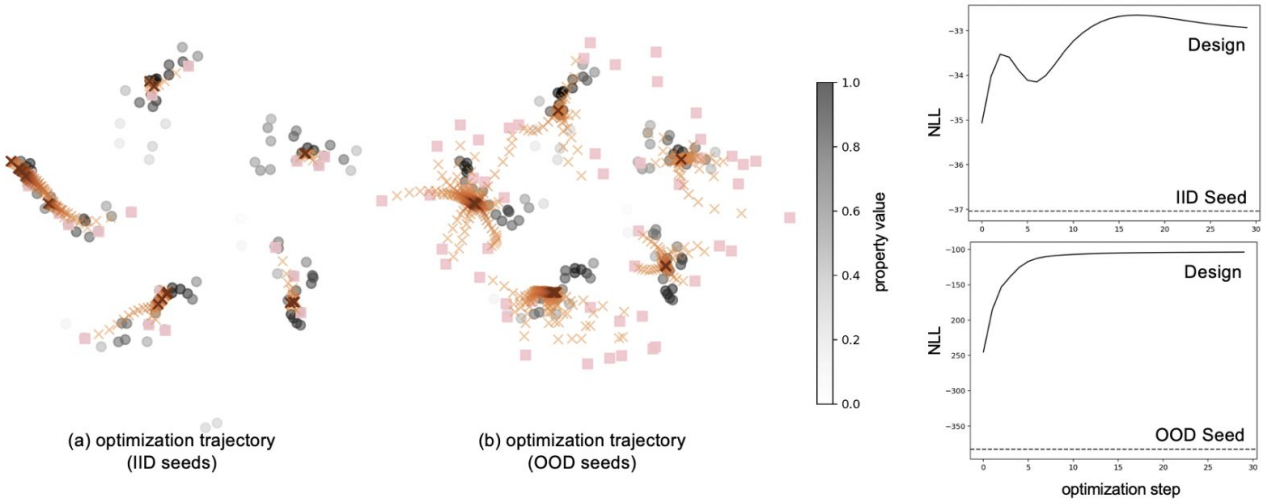
Theorem 1.
Let f^* be the optimal solution of the matched reconstruction objective with a sufficiently small Δ_x. For any point x in the matched dataset for which p is uniform within a ball of radius Δ_x, we have $f^*(x) \rightarrow c\nabla_g(x)$ for some positive constant c.

Property Enhancer - PropEn

- **Step 3: Iterative optimization and sampling**

Prescient Design
A Genentech Accelerator

At test time, we feed a seed design x_0 to PropEn, and read out an optimized design x_1 from the its output. We then proceed to iteratively re-feed the current design to PropEn until $f_\theta(x_t) = x_t$



(a) optimization trajectory (IID seeds)

(b) optimization trajectory (OOD seeds)

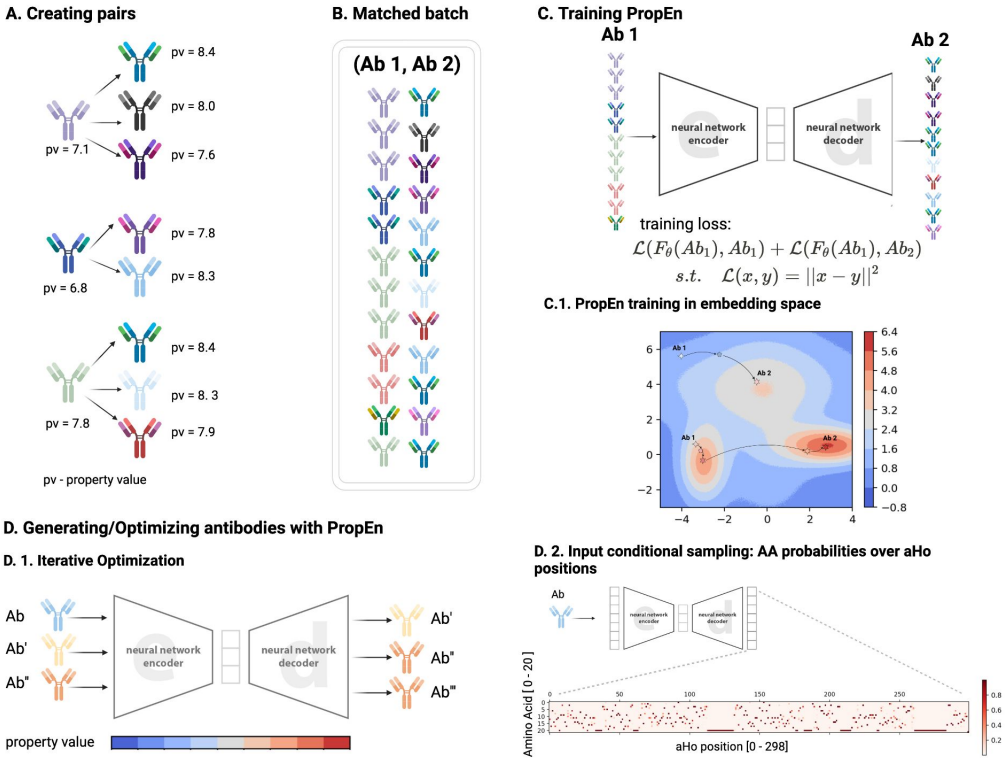
(c) NLL during optimization

Genentech
A Member of the Roche Group

In-vitro experiment: therapeutic protein optimization

Prescient Design
A Genentech Accelerator

Expression rate: ~95%
Binding rate: ~90%



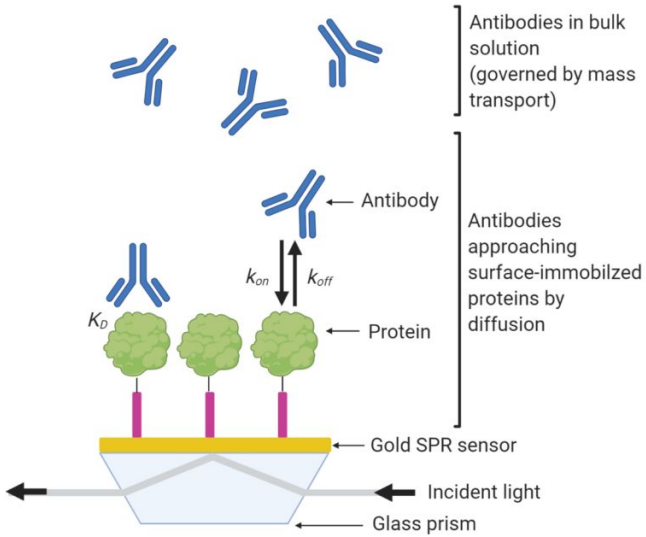
Genentech
A Member of the Roche Group

In-vitro experiment: therapeutic protein optimization

Prescient Design
A Genentech Accelerator

Experimental design

- **task**: optimizing the binding affinity of lead antibody molecule
- **metric**: negative log ratio of the association and dissociation constants (pKD)
- **data collection**: low-throughput Surface Plasmon Resonance (SPR) experiments
- **data**: 3 targets, 9 seeds



Baselines

Latent Diffusion - unguided and guided[3], discrete Walk-Jump Sampling [1], Lambo 2 [2]

[1] Frey, Nathan C., et al. "Protein discovery with discrete walk-jump sampling." *ICLR* 2024.
[2] Gruver, Nate, et al. "Protein design with guided discrete diffusion." *NeurIPS 2023*.
[3] Choi, J., et al. "Conditioning method for denoising diffusion probabilistic models. ICCV 2021.

Results - 1

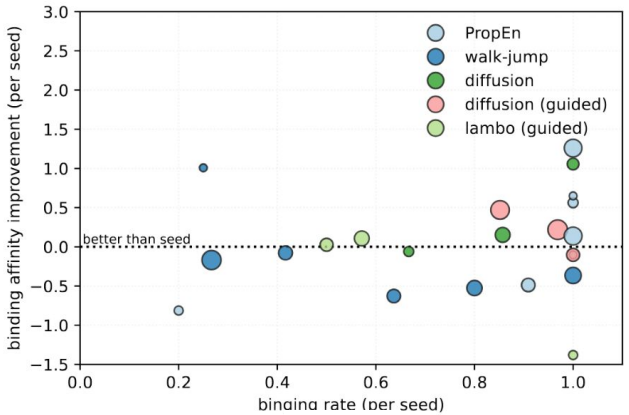
Table 1: Binding rate (and number of designs submitted). Higher is better.

	Herceptin	T1S1	T1S2	T1S3	T2S1	T2S2	T2S3	T2S4	overall
PropEn	90.9% (11)	100.0% (4)	100.0% (6)	100.0% (24)	20.0% (5)	100.0% (23)	100.0% (16)	100.0% (4)	**94.6% (93)**
walk-jump [7]	-	25.0% (4)	80.0% (15)	100.0% (18)	26.7% (30)	41.7% (12)	100.0% (15)	63.6% (11)	62.9% (105)
lambo (guided) [8]	50.0% (10)	0.0% (4)	-	100.0% (5)	0.0% (9)	-	100.0% (1)	57.1% (14)	44.2% (43)
diffusion	-	100.0% (8)	85.7% (14)	-	-	-	88.2% (17)	66.7% (6)	86.7% (45)
diffusion (guided)	-	85.2% (27)	96.9% (32)	-	-	-	93.3% (15)	100.0% (10)	92.9% (84)

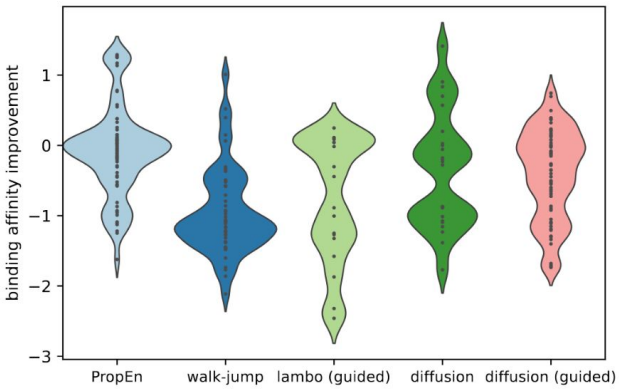
Table 2: Fraction of designs improving the seed and total designs tested. Higher is better.

	Herceptin	T1S1	T1S2	T1S3	T2S1	T2S2	T2S3	T2S4	overall
PropEn	0.0% (11)	100.0% (4)	33.3% (6)	41.7% (24)	0.0% (5)	69.6% (23)	0.0% (16)	0.0% (4)	**34.4% (93)**
walk-jump [7]	-	25.0% (4)	6.7% (15)	5.6% (18)	3.3% (30)	8.3% (12)	0.0% (15)	0.0% (11)	4.8% (105)
lambo (guided) [8]	10.0% (10)	0.0% (4)	-	0.0% (5)	0.0% (9)	-	0.0% (1)	35.7% (14)	14.0% (43)
diffusion	-	62.5% (8)	14.3% (14)	-	-	-	0.0% (17)	0.0% (6)	15.6% (45)
diffusion (guided)	-	51.9% (27)	15.6% (32)	-	-	-	0.0% (15)	0.0% (10)	22.6% (84)

Results - 2



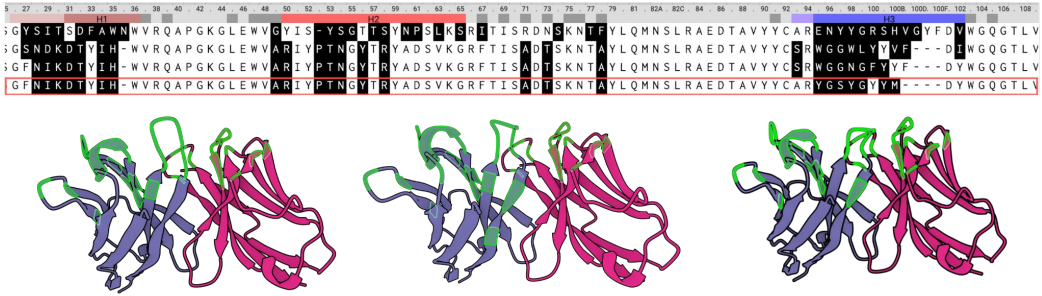
(a)



(b)

Figure 5: Therapeutic protein optimization results: (a) The left figure contrasts the binding rate with the 90-th percentile of the binding affinity improvement for each method and seed. Points on the top-right are on the Pareto front. (b) The right figure focuses on binders and reports the histograms of binding affinity improvement across all designs and seeds.

Summary & Outlook



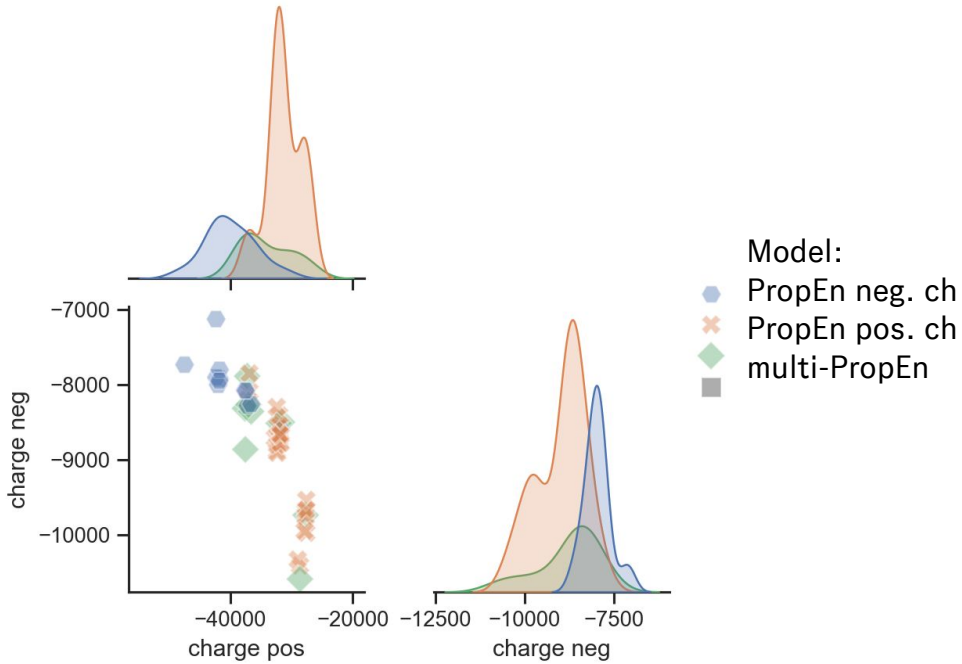
- ➔ property enhancement method <u>without discriminator</u> for a single or multiple properties
- ➔ data (modality) agnostic
- ➔ works well even in small - medium data regimes
- ➔ easy to train - no hyperparameter tuning

Bonus slides

Multi-property enhancer

- **Instead of single property, we can optimize for a multivariate score of a molecule**

Step 1: compute multivariate rank/score for multiple properties
Step 2: match and optimize designs for the multivariate score with Propen



Variations of PropEn

(PropEn) mix

- reconstruct both better design and the original

$$\ell(f, \hat{p}) = \mathbb{E}_{x \sim \hat{p}}[\mathbb{E}_{x' \sim \hat{\mu}_x}[\ell(x', f(x)) + \beta\, \ell(x, f(x))]]$$

- lets us stay close to the seed
- increases diversity

(PropEn) x2x reconstruct only the design

xy2xy reconstruct the design and the property value;

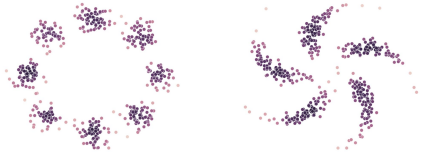
- helps stabilizing training
- allows for controlled generation

Variations of PropEn

- **ablation study on toy data**

Prescient Design
A Genentech Accelerator

(PropEn) mix reconstruct both better design and the original
(PropEn) x2x - reconstruct only the design
xy2xy - reconstruct the design and the property value;



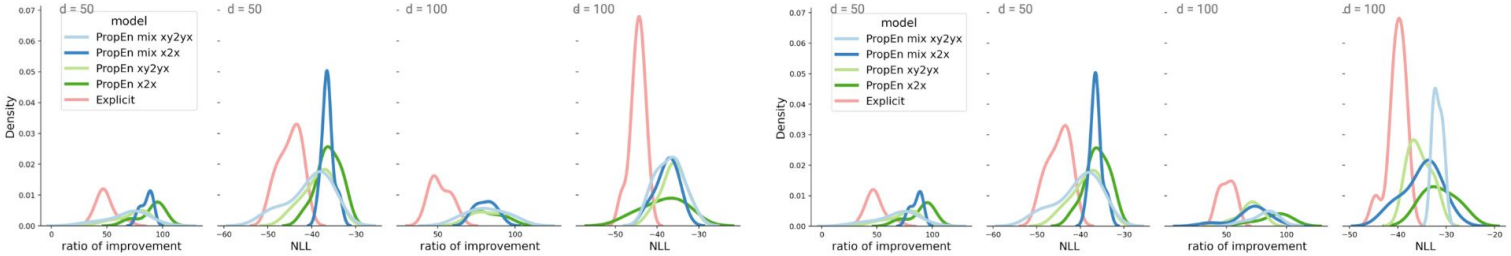


Figure 3: PropEn in toy examples in $d \in \{50, 100\}$, left side: 8-Gaussians, right side: pinwheel. Distribution of evaluation metrics from 10 repetitions of each experiment.

Genentech
A Member of the Roche Group