Implicit Guidance with PropEn: Match your data to follow the gradient

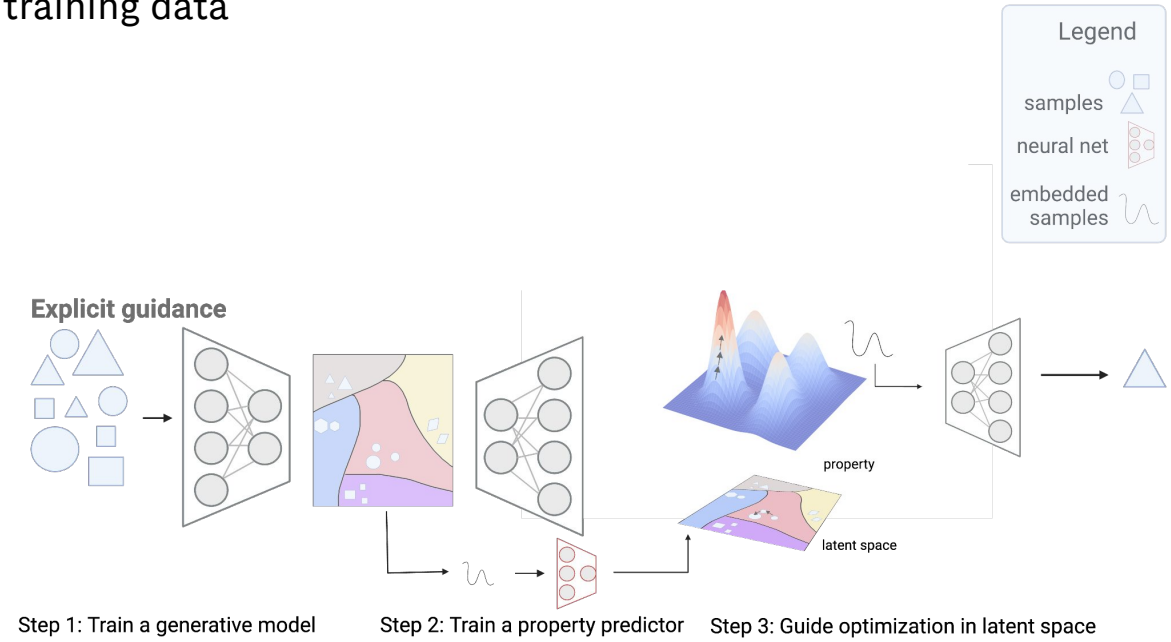
Natasa Tagasovska, Vlad Gligorijevic,
Kyunghyun Cho, Andreas Loukas

Genentech
A Member of the Roche Group

Implicit vs Explicit guidance

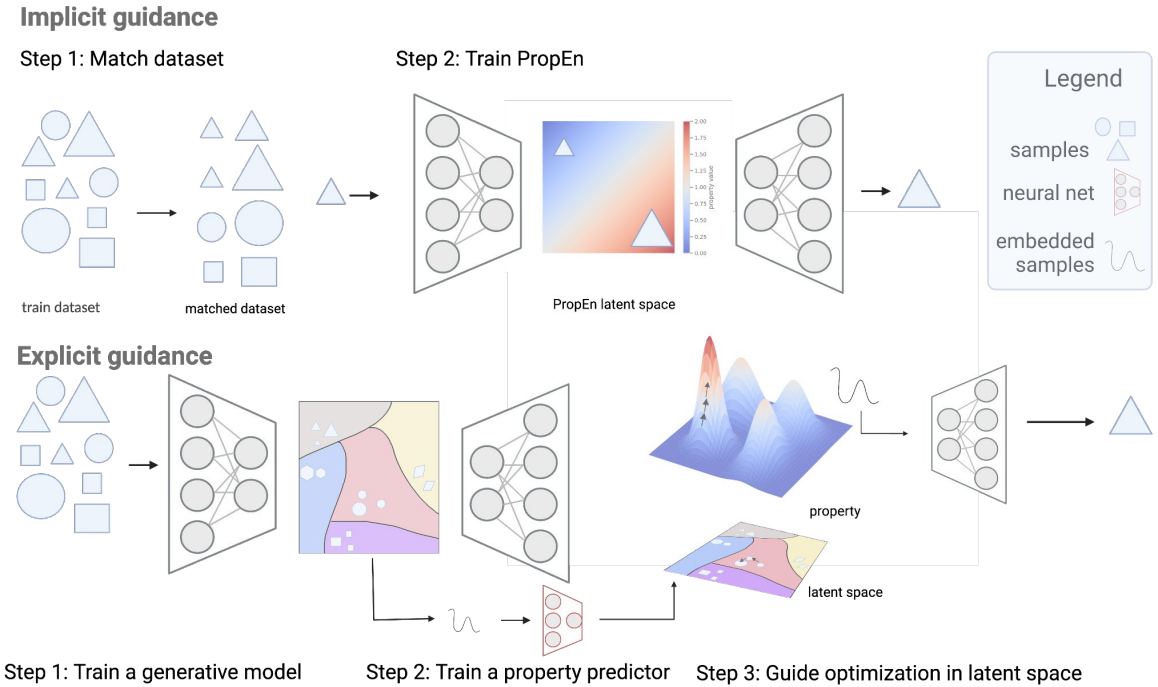
Explicit guidance requires:

- both a generative and discriminative model
- lot of training data



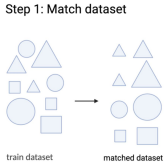
Implicit vs Explicit guidance

Implicit guidance doesn't require training a discriminative model and works even in small datasets!



Tagasovska, N., Gligorijević, V., Cho, K., & Loukas, A. (2024). Implicitly Guided Design with PropEn: Match your Data to Follow the Gradient. *arXiv preprint arXiv:2405.18075*.

Step 1: Match the 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.

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

Example: x - coordinates of polygon, y - area of shape, dist: Euclidian
x - antibody sequence, y - binding affinity, dist: edit/Levenstein
x - portfolio of stocks, y - portfolio value/risk, dist: Jaccard

Genentech
A Member of the Roche Group

Step 2: Approximate the gradient

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

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

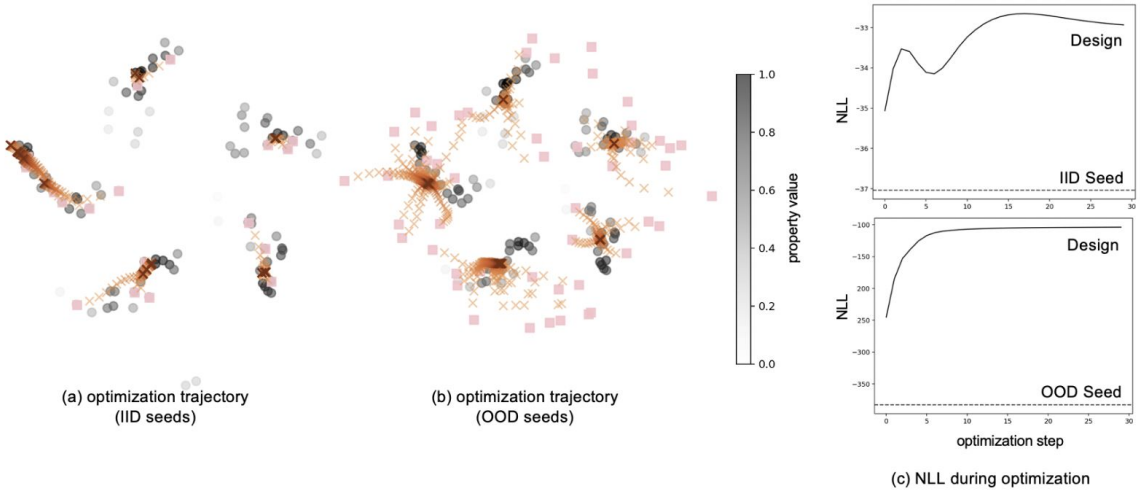
Where ℓ is an appropriate loss for the data in question, such as an mean-squared error (MSE) or cross-entropy loss.

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 $f^*(x) \to c \nabla_g(x)$ for some positive constant c.

Step 3: Optimize designs with implicit guidance

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 $f_\theta(x_t) = x_t$



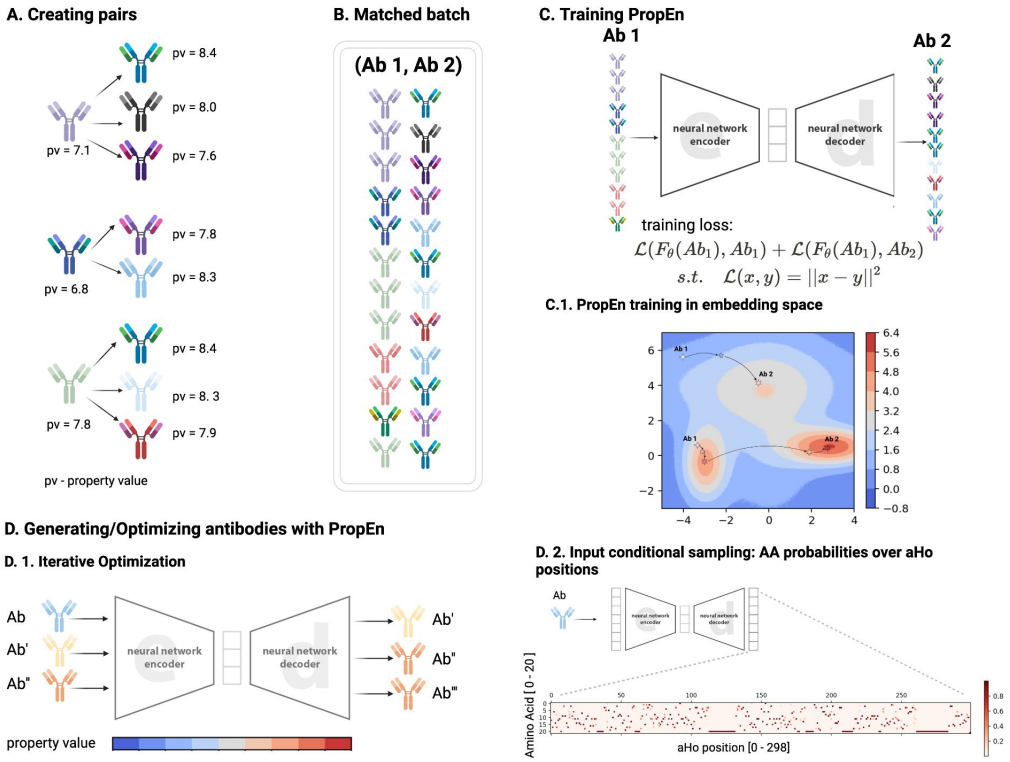
(a) optimization trajectory (IID seeds)

(b) optimization trajectory (OOD seeds)

(c) NLL during optimization

Genentech
A Member of the Roche Group

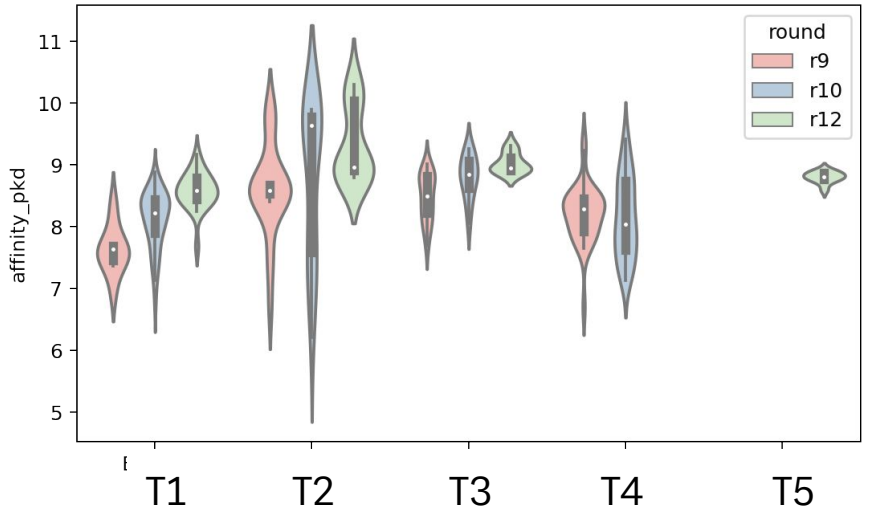
PropEn for Antibodies



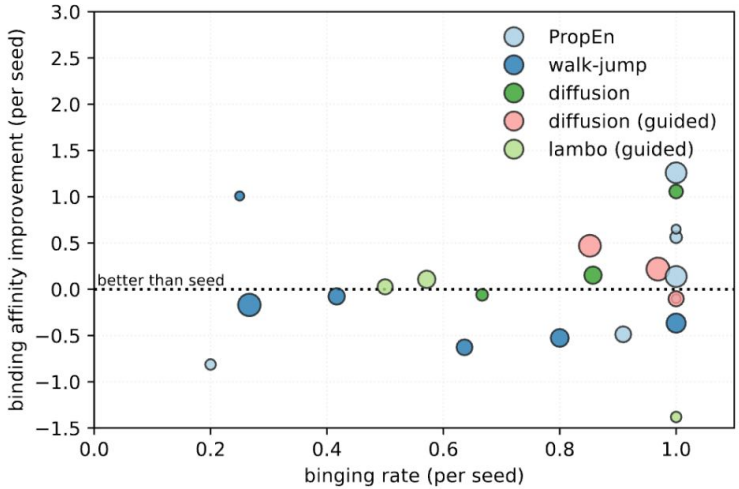
PropEn in LitL

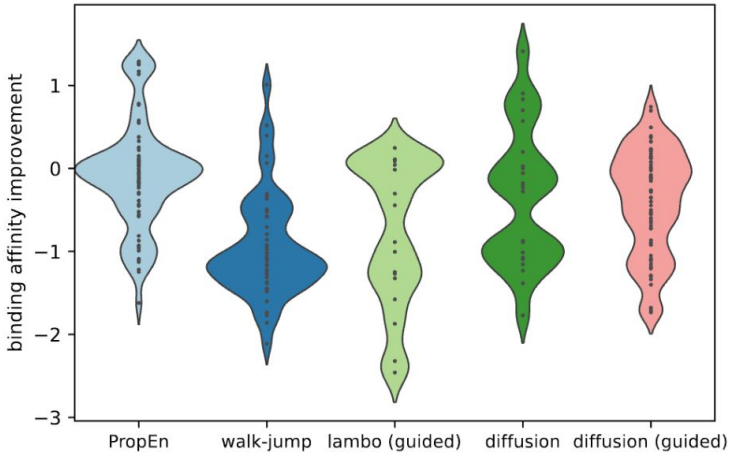
Expression rate: ~95%
Binding rate: ~90%

	Round 9	Round 10	Round 12
1x better binders	45/129 (34.9%)	98/247 (40%)	47/55 (85.45%)
3x better binders	12/129 (10%)	36/247 (15%)	31/55 (56.36%)
Highest improvement	5.6 (x seed)	32.8 (x seed)	38.1 (x seed)



PropEn vs baselines R9





Variations of PropEn

Genentech
A Member of the Roche Group

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;

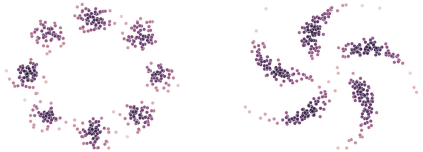
- helps stabilizing training
- allows for controlled generation

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;



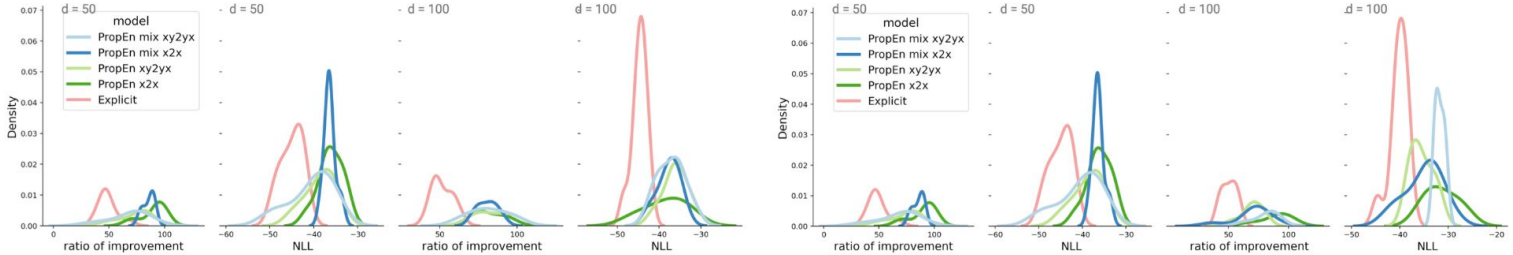


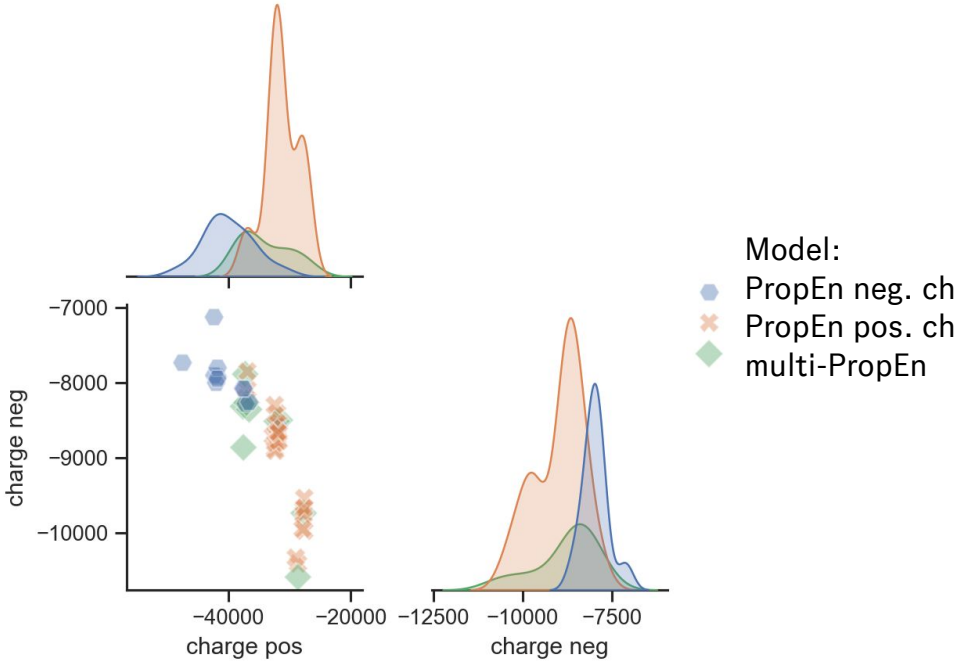
Figure 3: PropEn in toy examples in $d \in \{50, 100\}$, left side: 8-Gaussians, right side: pinwheel. Distribution of evaluation metrics from 10 repetitions of each experiment.

Genentech
A Member of the Roche Group

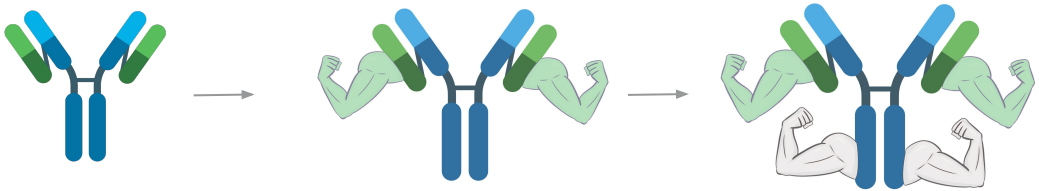
Multi-property enhancer

- **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



Summary and outlook



- ➔ property enhancement method without discriminator for a single or multiple properties
- ➔ data (modality) agnostic (see our preprint for example in aerodynamics engineering)
- ➔ works well even in small - medium data regimes
- ➔ easy to train - no hyperparameter tuning