Optimistic Games for Combinatorial Bayesian Optimization with Application to Protein Design

Melis Ilayda Bal · Pier Giuseppe Sessa · Mojmir Mutny · Andreas Krause

The Thirteenth International Conference on Learning Representations
ICLR 2025

Combinatorial Bayesian Optimization

$$x^* \in \arg\max_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X}$: combinatorial and unstructured

$$\{x^1, \ldots, x^n\}, x^i \in \mathcal{X}^{(i)}, \qquad \mathcal{X} = \prod_{i=1}^{n} \mathcal{X}^{(i)}$$

Example: Protein design

Q?F?DVF?GK??...E?PM?A → 20^{300} possibilities

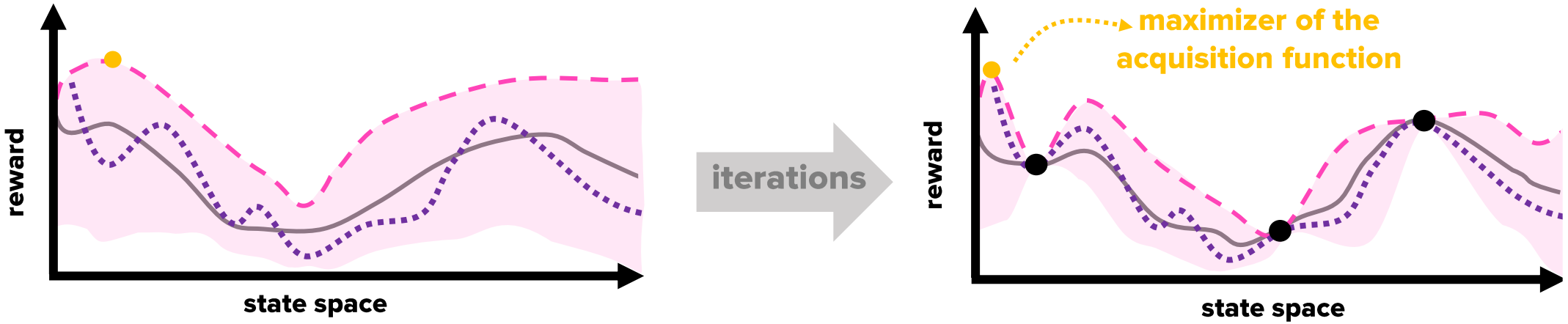
300 amino acid sites

GFP protein

Combinatorial Bayesian Optimization

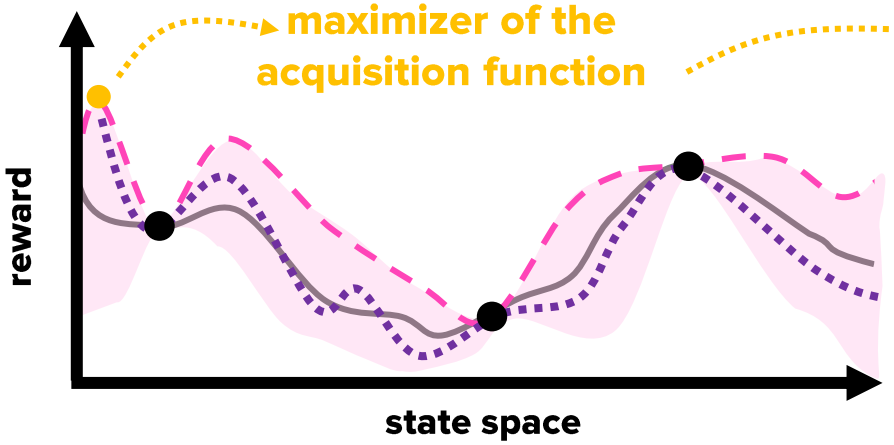
$$x^* \in \arg\max_{x \in \mathcal{X}} f(x),$$

where $\mathcal{X}$: combinatorial and unstructured



acquisition function
model fit
uncertainty
true function
measurements

Combinatorial Bayesian Optimization



Q?F?DVF?GK??...E?PM?A

20^{300} possibilities

intractable

GameOpt: Optimistic Games

GameOpt: Optimistic Games

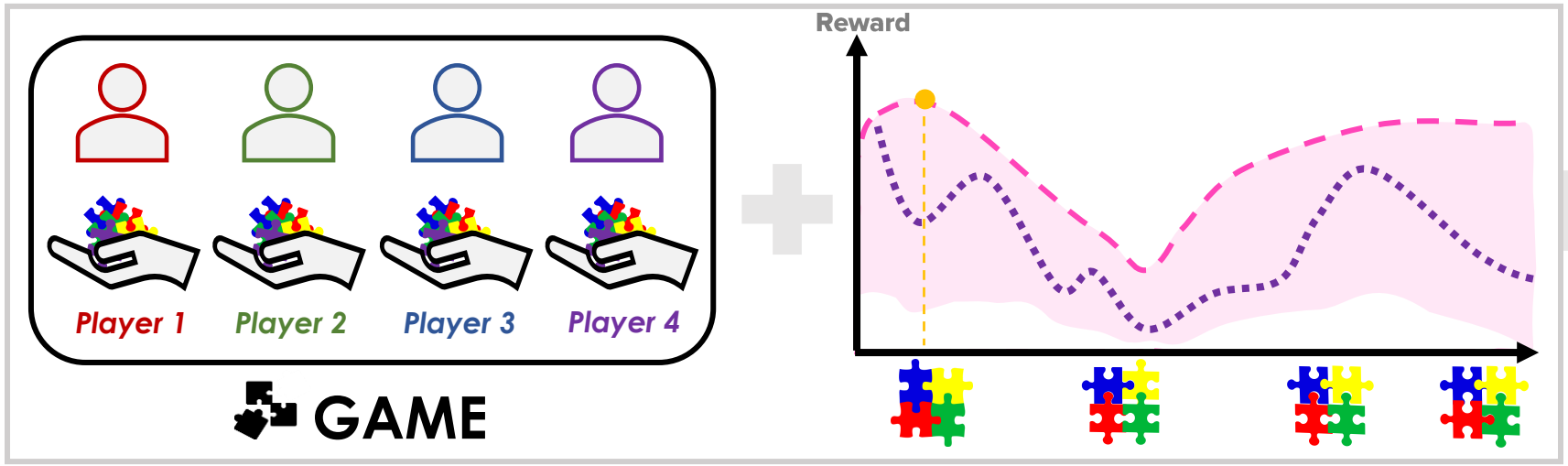
Optimistic Games - Idea

Cooperative game

Game equilibria

Maximization of acquisition function

Evaluation points



EQUILIBRIUM

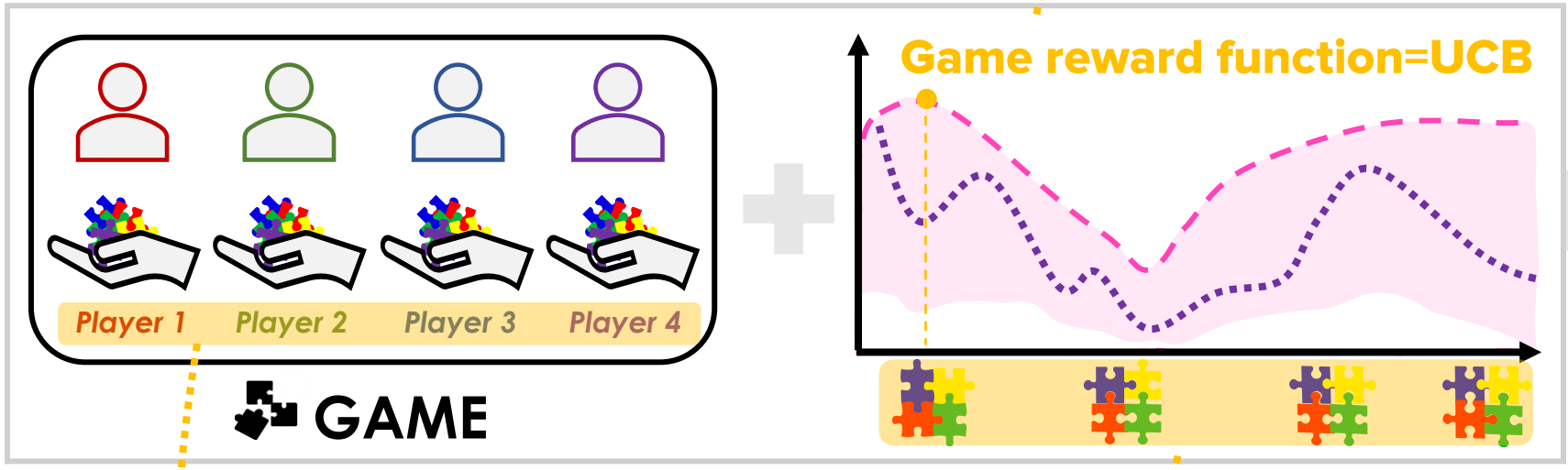
EQUILIBRIUM

Optimistic Games - Idea

$$x^* \in \arg\max_{x \in \mathcal{X}} f(x)$$

$$UCB(\mathcal{GP}, \cdot): \prod_{i=1}^{n} \mathcal{X}^{(i)} \to \mathbb{R}$$

$$UCB_t(\mathcal{GP}, x) = \mu_t(x) + \beta\sigma_t(x)$$



Candidate evaluation points

EQUILIBRIUM

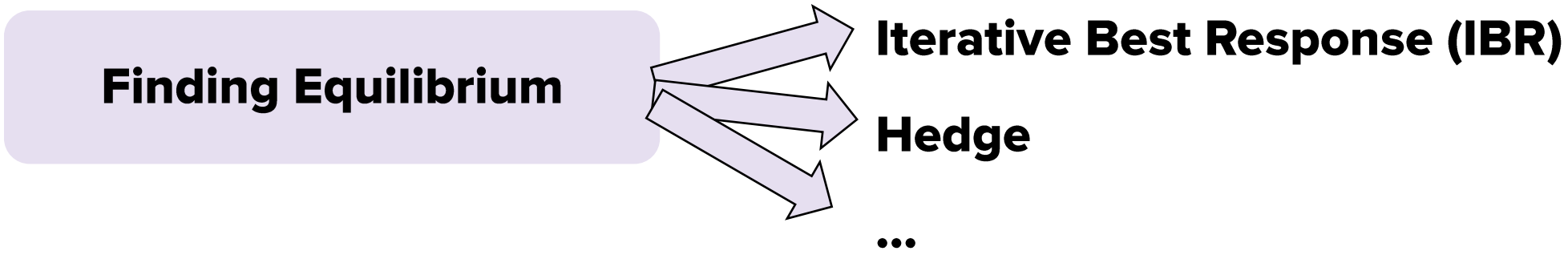
EQUILIBRIUM

$$\{x^1, \ldots, x^n\}, x^i \in \mathcal{X}^{(i)}$$

$$\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}^{(i)}$$

Optimistic Games - Equilibrium Finding

The goal of players → compute (Nash) equilibria



Application to Protein Design

Application to Protein Design

Protein: QCFFDVFQGKSRPMFA → amino acid sequences

- Protein length = L
- Possible protein configurations → 20^L

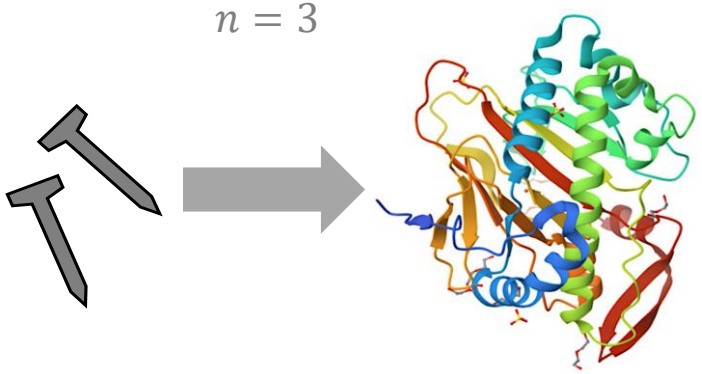
GFP protein

Goal: Find optimal amino acid sequence to maximize the functional capacity (fitness)

Application to Protein Design - Datasets

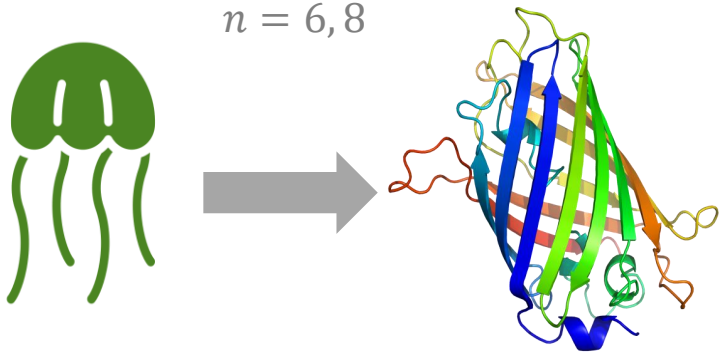
Halogenase[1], L=3

$n = 3$



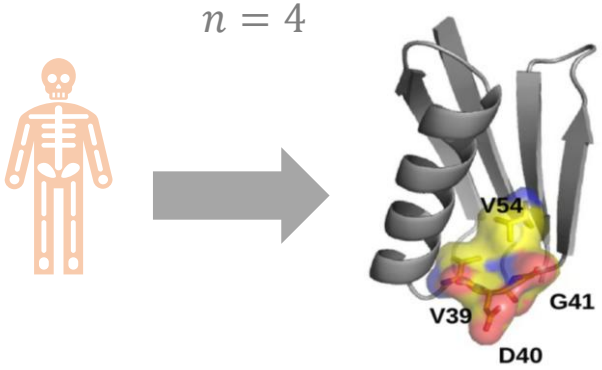
GFP[3], L=238

$n = 6,8$



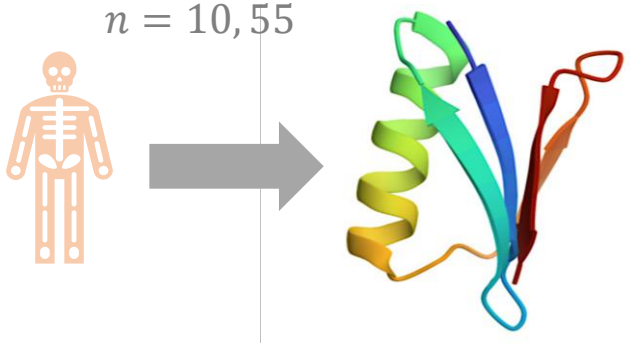
GB1(4)[2], L=4

$n = 4$



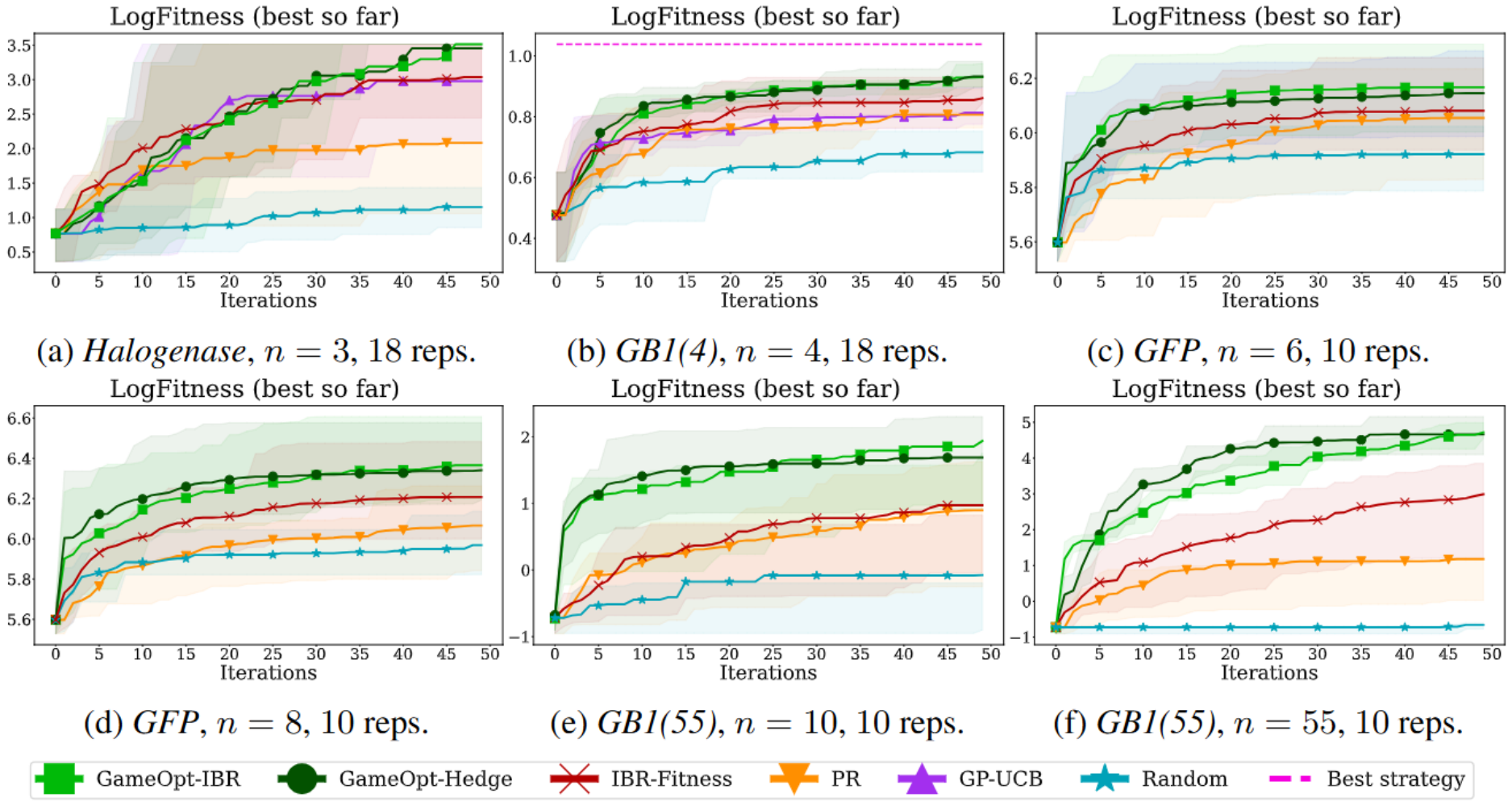
GB1(55)[4], L=55

$n = 10,55$

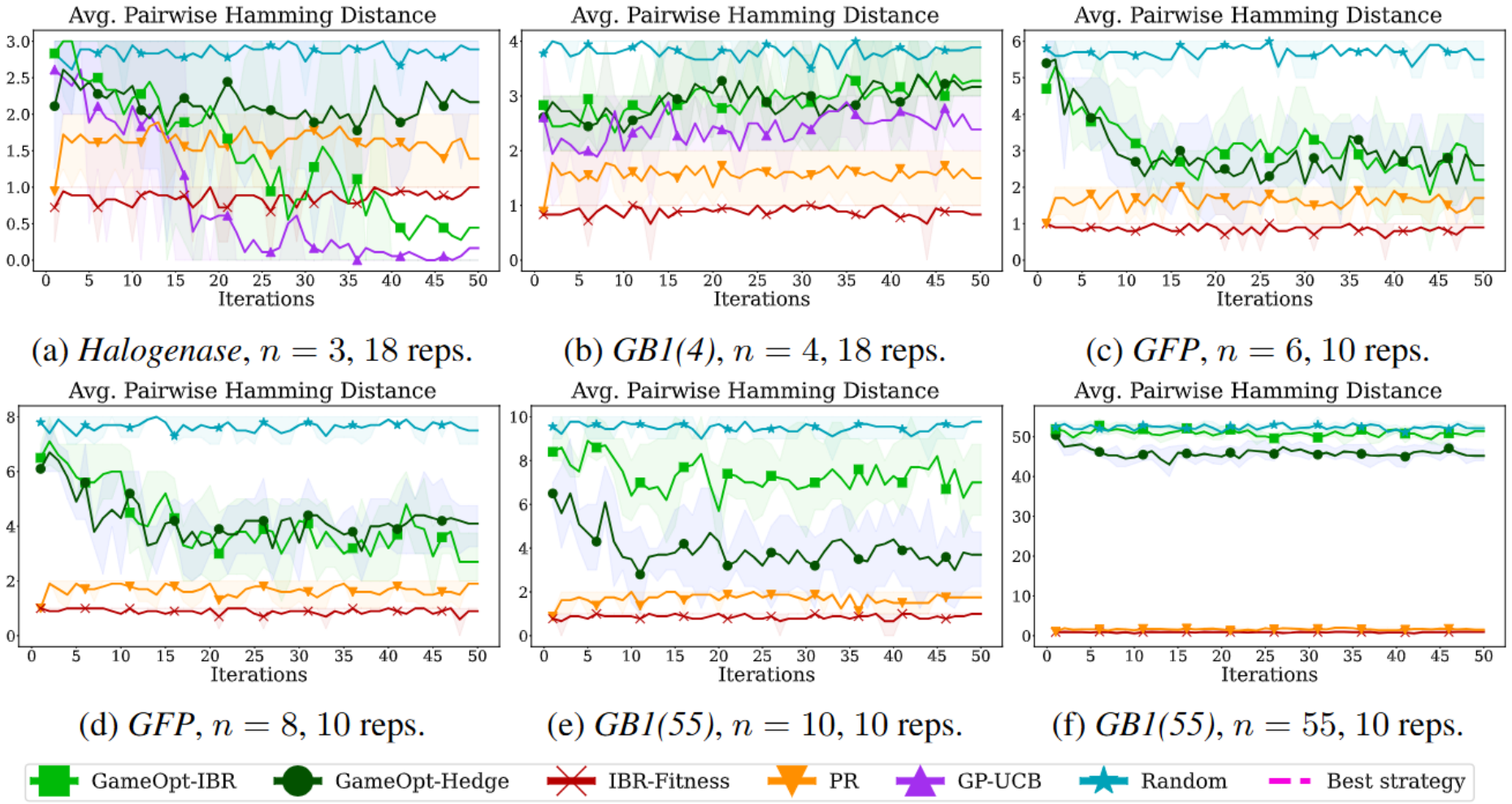


[1] Büchler, J., Malca, S. H., Patsch, D., Voss, M., Turner, N. J., Bornscheuer, U. T., Allemann, O., Le Chapelain, C., Lumbroso, A., Loiseleur, O., et al. Algorithm-aided engineering of aliphatic halogenase welo5* for the asymmetric late-stage functionalization of soraphens. Nature Communications, 2022.
[2] Wu, N. C., Dai, L., Olson, A., Lloyd-Smith, J. O., and Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife, 2016.
[3] Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. Nature methods, 2021.
[4] Olson, C. A., Wu, N. C., and Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. Current biology, 2014.

Application to Protein Design – Convergence Speed



Application to Protein Design – Pairwise distance



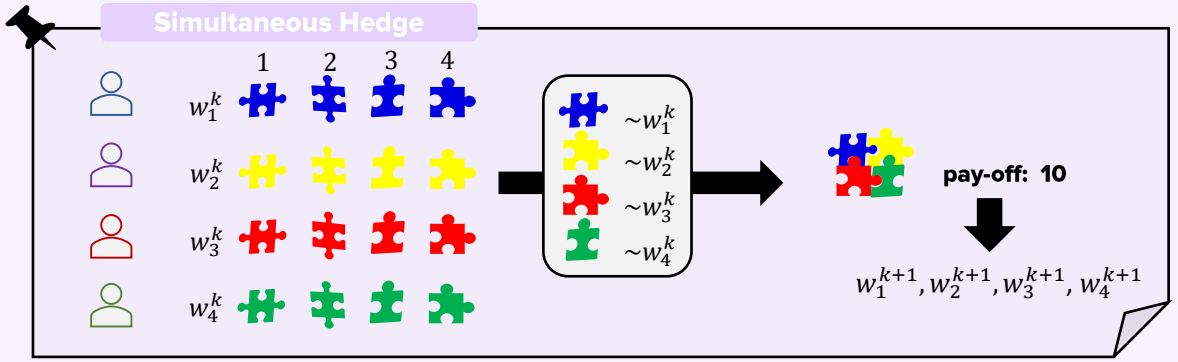
Further...

Sample complexity

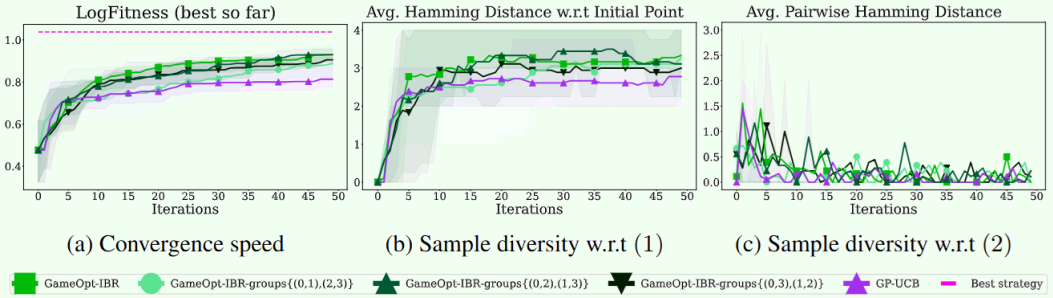
Theorem 4.2 (Sample complexity of GAMEOPT). *Assume f satisfies the regularity assumptions of Section 2, and* GAMEOPT *is run with confidence width $\beta_t = 2n \log\left(\sup_{i\in\mathcal{N}} |\mathcal{X}_i| \frac{t^2\pi^2}{6\delta}\right)$. Then, with probability at least $1-\delta$ and for a given accuracy $\epsilon \geq 0$, the strategy $x_{T^\star}$ returned by* GAMEOPT *is a ϵ-approximate Nash equilibrium when*

$$T \geq \Omega\left(\frac{\beta_T \gamma_T}{\epsilon^2}\right). \tag{4}$$

Equilibrium finding subroutines



Player grouping



(a) Convergence speed (b) Sample diversity w.r.t (1) (c) Sample diversity w.r.t (2)

Optimistic Games for Combinatorial Bayesian Optimization with Application to Protein Design

ICLR poster:
Thu 24 Apr 15:00, Poster Session #2

ArXiv

Code