enzyan BIOCATALYSIS
WHITEPAPER

c.ascend: Bayesian Optimization of Biocatalytic Cascades

Ascending the fitness peak of a multi-enzyme cascade

Optimizing a multienzyme biocatalytic cascade for both yield and cost is a multi-objective design problem in a search space too large for trial-and-error or classical Design of Experiments. We present c.ascend: a multi-objective Bayesian optimization workflow tailored to biocatalytic cascade optimization. The method learns from each round of experiments and proposes the next batch of reactions to maximally extend the set of attainable yield-cost trade-offs. Here we investigated a four-enzyme, one-pot cascade for the synthesis of an API as case study. Within just two weeks and four rounds of 24 reactions, the yield could be increased from 66% to 82% while biocatalyst costs could be reduced by 2590 €/kg product. The optimizer converged on a Pareto front of seven non-dominated operating points and the strategically collected dataset informs decision making for enzyme engineering and downstream processing. The workflow applies to any cascade where reactions are expensive, batch turnaround is days, and multiple competing objectives must be traded off against each other.

c.ascend
BIOCATALYTIC CASCADE OPTIMIZER

Multi-enzyme cascades enable the synthesis of complex molecules in a single reaction vessel,[1,2] but their optimization is challenging because many interacting components must be balanced precisely to maintain mass flux, avoid side reactions, and limit intermediate degradation. A four-enzyme system with ten possible concentration levels per enzyme already generates 10,000 possible recipes. This number grows exponentially as concentration resolution or system complexity increases. Exhaustively testing such spaces is impractical, while trial-and-error is highly sample-inefficient and unlikely to identify truly optimal conditions within a realistic timeframe. Kinetic modelling is one possible route to optimization because it reflects the physical behavior of the system and can, in principle, predict favorable reaction conditions.[3] However, it depends on detailed mechanistic knowledge and reliable kinetic parameters for all enzymes involved, which are often difficult or impossible to obtain; especially when unstable intermediates are present or operating conditions differ strongly from assay conditions.[4] Classical Design of Experiments (DoE) offers a more empirical alternative, but it is best suited to relatively simple, near-quadratic response surfaces and single-optimum objectives.[5] In enzyme cascades, by contrast, substrate inhibition, cofactor imbalance, and cross-reactivity often create far more complex behaviors. In addition, cascade optimization is rarely a matter of maximizing yield alone. In practice, yield must be balanced against cost, since increasing enzyme concentrations often improves conversion only up to a point, after which additional enzyme becomes economically inefficient. As a result, there is typically no single best operating condition, but rather a range of optimal trade-offs depending on process priorities. These non-dominated solutions form the *Pareto front*, which captures the most relevant yield-cost compromises and provides a more realistic target for optimization than any single maximum.

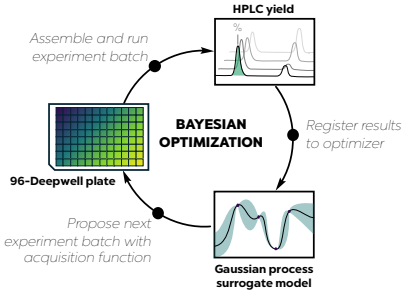
Here we introduce c.ascend: a Bayesian optimization-based workflow tailored to biocatalytic cascades. As a worked example, we use a three-step, one-pot cascade with four enzymes for the synthesis of an API. The methodology generalizes to any cascade or process where reactions are expensive, batches take days, and multiple objectives compete.

[1] Klapars, A. et al. Biocatalytic cascades enable manufacture of the macrocyclic peptide enlicitide. *Science* **2026**, *392*, 643-647. DOI: 10.1126/science.aed8713.
[2] Huffman, M.A. et al. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **2019**, *366*, 1255-1259. DOI: 10.1126/science.aay8484.
[3] Finnigan, W. et al. Engineering a Seven Enzyme Biotransformation using Mathematical Modelling and Characterized Enzyme Parts. *ChemCatChem* **2019**, *11*, 3474-3489. DOI: 10.1002/cctc.201900646.
[4] Kuschmierz, L. et al. Workflows for optimization of enzyme cascades and whole cell catalysis based on enzyme kinetic characterization and pathway modelling. *Curr. Opin. Biotechnol.* **2022**, *74*, 55-60. DOI: 10.1016/j.copbio.2021.10.020.
[5] Petermeier, P. et al. Design of a green chemoenzymatic cascade for scalable synthesis of bio-based styrene alternatives. *Green Chem.* **2022**, *24*, 6889-6899. DOI: 10.1039/d2gc01629j.
[6] Rasmussen, C. E., Williams, C. K. I. Gaussian Processes for Machine Learning. MIT Press (**2005**). DOI: 10.7551/mitpress/3206.001.0001.
[7] Balandat, M. et al. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. *Advances in Neural Information Processing Systems (NeurIPS)* **2020**, *33*. DOI: 10.48550/arXiv.1910.06403.
[8] Shahriari, B. et al. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **2016**, *104*, 148-175. DOI: 10.1109/JPROC.2015.2494218.
[9] Bakshy, E. et al. AE: A domain-agnostic platform for adaptive experimentation. NeurIPS Systems for ML Workshop **2018**. (Ax platform.)

A primer on Bayesian optimization

Bayesian optimization (BO) is a strategy for sample-efficient experimental design. After each round of measurements, BO fits a *surrogate model* – usually a Gaussian process (GP)[6] – that predicts the outcome at any untested combination of inputs and quantifies its uncertainty. The next batch of experiments is chosen to be maximally informative for the goal at hand: high model uncertainty in promising regions makes a candidate attractive (exploration); high predicted performance also makes it attractive (exploitation). The algorithm balances both automatically (Figure 1A).[7,8,9] BO is well-suited to problems where experiments are expensive, the response surface is unknown and likely complex, and the goal is to map a trade-off curve – a good fit with biocatalytic cascades.

A Concept of Bayesian optimization.



B Pareto-front and hypervolume in a two-objective space..

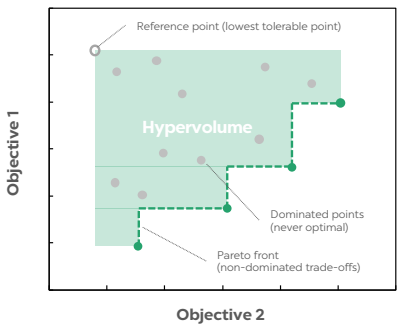


Figure 1. **A** The concept of iterative Bayesian optimization in c.ascend. **B** The Pareto front and hypervolume in two-objective space. Points on the front are non-dominated trade-offs; points behind the front are dominated and never optimal. The shaded region is the hypervolume: a single number that summarizes how much of the desirable objective space the front covers. Bayesian optimization proposes new experiments to extend this region.

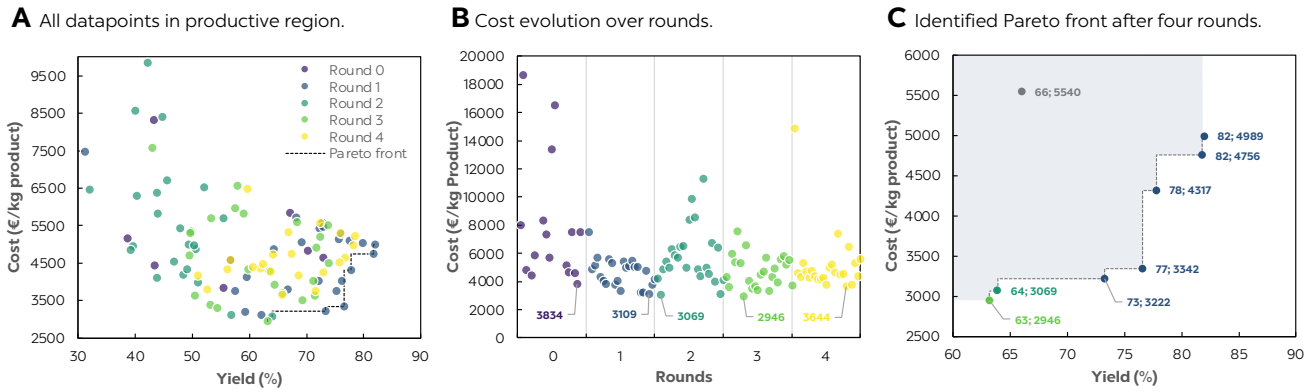


Figure 2. A Yield versus cost per kg product, with each round in a distinct color. The Pareto front evolved at the bottom right in the low cost-high yield region. **B** Per-round distribution of cost per kg product, with the lowest-cost reaction in each round annotated. The lowest-cost data point captures the algorithm's progress at the cheap end of the front, where small enzyme savings compound into large cost reductions. Note that the cheapest conditions not always are Pareto-optimal. **C** The observed Pareto front as dotted line. The x-axis is clipped at 60% yield to improve visibility. The non-dominated points are labelled with their corresponding yield and cost colored according to the round in which they were discovered. The benchmark conditions prior to optimization are shown as grey dot.

For the multi-objective optimization problems in c.ascend we use *qLogNEHVI* (q-batch Logarithmic Noisy Expected Hypervolume Improvement) as acquisition function. The figure of merit it maximizes is the *hypervolume* of the region of objective space dominated by the current Pareto front, measured from a reference point (Figure 1B). Each new batch is chosen to maximally grow this hypervolume in expectation.[10,11]

Search-space definition and seed round

A preliminary manual optimization led to 66% product yield with a total enzyme loading of 27.5 g/L, corresponding to 5540 € material costs per kg product. This represents the benchmark of the c.ascend campaign. The buffer composition and reagent concentration were also pre-optimized and kept constant in this example. The design space is thus limited to the four enzyme concentrations only, although e.g. co-factor/co-substrate concentration, ionic strength or pH-value could readily be included in the search space as well.

The c.ascend workflow begins with a Round 0 of 24 reactions randomly chosen from a Latin hypercube sampling (LHS)[12] cohort, which is a space-filling design that distributes points evenly across the 4-dimensional search space. No Bayesian reasoning is involved yet. This initial round is used to seed the surrogate model with broad, unbiased information and make the experiment batch suggestion for Round 1.

The final Pareto front and what it means

Across four rounds and two weeks of experimentation, c.ascend efficiently mapped the yield–cost trade-off in the enzyme cascade by alternating exploitation of promising regions with targeted exploration of uncertain or potentially cheaper conditions (Figure 2A&B). The hypervolume plateau between Rounds 3 and 4 confirmed convergence and justified stopping the campaign. *The shape of the resulting seven-point Pareto front carries process information* as each point is a viable operating recipe; the choice between them depends on what the downstream process prioritizes. For example, there is a high-yield point (82%, 4990 €/kg) where downstream purification benefits from clean conversion and a lowest-cost point (63%, 2950 €/kg) for cost-driven processes that can absorb a longer purification stage. Across the front, increasing yield from 63% to 82% costs roughly 2100 €/kg: a 65% premium for a 30% relative yield gain. Whether that premium is worthwhile depends on downstream economics that are external to the optimization itself. The value of the front is that this question becomes answerable rather than guessed at.

Compared to the benchmark conditions prior to the optimization, *the yield could be increased by 16%* while *costs could be reduced by 2590 €/kg product*. Interestingly, most of the Pareto-optimal points in the high-yield region were identified already in the first round (i.e. after 24 reactions), highlighting the viability of the chosen acquisition function.

Statistical analysis of the strategically collected dataset further allowed to identify the bottlenecks of the cascade and allocate engineering resources where they are most needed (Table 1). E3 was found to have the strongest correlation with yield and was assigned the highest priority for engineering. In line with experimental observations, a negative correlation of E1 with yield hints at formation of an unstable intermediate which should be consumed as fast as possible. Indeed, E2 is ranked second in the engineering priority to improve mass flux.

Table 1. Enzyme influence on biocatalytic yield — statistical summary.

Enzyme	Partial r[a]	Relative Feature Importance[b]	Effect Direction	Engineering Priority
E3	+0.55	0.54±0.07	Positive	1st
E2	+0.32	0.54±0.08	Positive	2nd
E4	+0.35	0.44±0.06	Positive	3rd
E1	−0.27	0.23±0.03	Negative	–

(**a**) Spearman partial correlation coefficient between each enzyme and yield, computed after removing the linear contribution of all other three enzymes by ordinary least-squares regression. (**b**) Relative Feature Importance: mean permutation importance from a Random Forest regression model (500 trees, 100 permutation repeats, $R^2 = 0.96$) trained on all four enzyme amounts to predict yield. For each enzyme, its values in the test set were randomly shuffled and the resulting drop in model accuracy (mean squared error) was recorded; larger values indicate greater influence on the predicted yield.

Conclusions

Our multi-objective Bayesian optimization workflow c.ascend is practical for biocatalytic cascade and process development. It scales beyond what classical Design of Experiments can handle and is easier to execute than kinetic modelling. It produces a Pareto front of operating points rather than a single optimum, and provides an internal stopping signal via hypervolume plateau, so resources are not wasted on rounds that cannot improve the result. The case study presented here balanced four enzymes using data from 96 reactions over four rounds (two weeks of experimental work), but the framework applies unchanged to cascades with more components, additional process variables (pH, salinity, temperature, cofactor concentrations, cosolvents etc.), or other combinations of objectives such as space-time yield, E-factor, or environmental metrics. The strategically collected data also provides a empirical basis for further development decisions such as enzyme engineering.

[10] Daulton, S.; Balandat, M.; Bakshy, E. Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. *Advances in Neural Information Processing Systems (NeurIPS)* **2021**, *34*. DOI: 10.48550/arXiv.2015.08195.
[11] Ament, S. et al. Unexpected Improvements to Expected Improvement for Bayesian Optimization. *Advances in Neural Information Processing Systems (NeurIPS)* **2023**, *36*. DOI: 10.48550/arXiv.2310.20708.
[12] Shields, M.D. et al The generalization of Latin hypercube sampling. arXiv:1507.06716v1.