Efficient Cell Factory Design by Combining Meta-Heuristic Algorithm with Enzyme Constrained Metabolic Models

Wenbin Liao[1,5], Gengrong Gao[2,5], Xingcun Fan[1,2], Haoyu Wang[2,3,4], Siwei He[2], Guangming Xiang[2], Xuefeng Yan[1]*, Hongzhong Lu[2]*

[1]Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, P. R. China

[2]State Key Laboratory of Microbial Metabolism, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China

[3]Division of Biotechnology, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, 116023, P. R. China

[4]University of Chinese Academy of Sciences, Beijing, 100049, P. R. China

[5]These authors contributed equally.

*Corresponding authors.

Correspondence: xfyan@ecust.edu.cn and hongzhonglu@sjtu.edu.cn

Abstract

The design of high-performance microbial cell factories is essential for advancing sustainable biomanufacturing. However, the intricate nature of metabolic networks complicates the prediction of genetic interventions, making strain optimization a challenging combinatorial problem. Here we present a novel computational strain design algorithm-MetaStrain that integrates enzyme-constrained models (ecModels) with efficient meta-heuristic algorithms to identify non-intuitive combinatorial gene targets for improving product yields. Our algorithm employs a modified enforced objective flux scanning algorithm to reduce the dimensionality of gene candidates and annotate editing strategies. Subsequently, binary and symbolic genetic algorithms and the adaptive differential evolution algorithm with external archives (JADE) are employed for gene target selection, with product yield as the fitness function. We adapted minimization of metabolic adjustment strategy for ecModels to evaluate the productive phenotypes of mutant strains. Computational validation using *Saccharomyces cerevisiae* as a host demonstrates significant enhancements in 2-phenylethanol production. Further generalization to spermidine biosynthesis confirms the framework's robustness. The proposed algorithm achieves high computational efficiency, stable convergence, and strong adaptability across diverse metabolic targets. This study provides a

powerful computational tool for metabolic engineering, addressing the complexity of strain design and highlighting the potential of meta-heuristic optimization in synthetic biology.

Key words: enzyme-constraint model, meta-heuristic algorithm, minimization of metabolic adjustment, *Saccharomyces cerevisiae*, strain design

Introduction

In the 21st century, the rapid advancements in Artificial Intelligence (AI) and synthetic biology[1] have significantly expanded the applications of microorganisms across diverse fields, including biofuels[2, 3], pharmaceuticals[4], food additives[5], and industrial enzymes[6]. The increasing global demand for petroleum-based products, coupled with concerns about environmental sustainability, has driven a transition toward microbial fermentation-based bioproduction pathways as an alternative to traditional chemical synthesis[7]. However, succeeding in the competitive landscape of traditional chemical production requires the development of high-performance Microbial Cell Factories (MCFs)[8]. This necessitates the reprogramming of cells through metabolic engineering to create optimal strains with superior production metrics, including titer, rate, and yield (TRY)[9].

Traditional strain optimization approaches rely heavily on researchers' expertise, often guided by qualitative and intuitive design principles, requiring extensive trial-and-error experimentation. This process is both time-consuming, costly and limited in scope. To address these challenges, constraint-based modeling[10] has facilitated the development of reliable quantitative metabolic models, providing a global perspective and a platform for system-level analysis in metabolic engineering. Genome-scale metabolic models (GEMs) provide a mathematical representation of metabolic networks, supporting pathway analysis, metabolic flux prediction, and strain design. GEMs can also integrate multi-omics data to improve predictive accuracy[11]. To enhance model-based in silico strain design, the GEMs were further enhanced with enzymatic constraints using kinetic and omics data[12]. This approach significantly reduces flux variability in traditional GEMs while revealing enzyme usage and distribution across metabolic pathways. It also overcomes the limitations of traditional GEMs in simulating metabolic shifts, such as overflow metabolism[13], the Crabtree effect [14], and the Warburg effect[15]. The vast number of metabolites and reactions in these models results in high-dimensional solution spaces. Flux balance analysis (FBA)[16] is widely used to simulate metabolic phenotypes. Additionally, the minimization of metabolic adjustment (MOMA)[17] was developed to simulate metabolic flux distribution in mutant strains following genetic interventions. Both FBA and MOMA have been extensively utilized for predicting gene targets in various computational strain design algorithms, but MOMA needs to be adjusted to ecModels.

A range of computational tools has been developed for *in silico* strain design, yet many face significant limitations. Early methods, such as OptKnock[18], employed bilevel optimization to identify gene deletion strategies, successfully increasing succinic acid production in *Escherichia coli*. OptGene[19] pioneered the use of evolutionary algorithms for identifying deletion targets, optimizing production of compounds like succinic acid, glycerol, and vanillin in *S. cerevisiae*. However, these tools were limited to knockout strategies. To address over-expression targets, FSEOF[20] applies additional constraints on product formation fluxes, identifying genes with increased flux contributions. This method has been successfully used to improve lycopene production but lacks the ability to explore combinatorial editing strategies. More advanced tools, such as OptCouple[21], extend the scope to include combinations of knockouts, insertions, and medium supplements, enabling growth-coupled production of target compounds. Similarly, OptForce[22] and OptDesign[23] rely on alternative biological objectives, analyzing flux distribution differences between wild-type and production strains to propose minimal intervention strategies. However, these tools largely rely on traditional GEMs, limiting their biological reliability. ecFactory[24] successfully predicts engineering targets for 103 valuable chemicals using the ecModel of yeast, but it provides a large number of targets for each product. Incorporating advanced AI techniques, such as reinforcement learning and machine learning, into strain design has also been explored[25, 26]. Furthermore, these tools often struggle with the combinatorial explosion inherent in large-scale metabolic networks, making it difficult to identify optimal solutions[27].

Meta-heuristic algorithms, as representatives of symbolic AI, are powerful tools for solving complex combinatorial optimization problems due to their strong search capabilities and adaptability. Genetic algorithms (GAs)[28], for instance, simulate natural selection and heredity to achieve rapid searches in high-dimensional spaces. Differential evolution (DE)[29], known for its simplicity and robustness, implement iterative updates through evolution operators, gradually approaching optimal solutions. Advanced strategies like elite retention and adaptive parameter control have significantly improved algorithmic performance[30, 31]. Meta-heuristic algorithms have shown great potential in addressing the complex optimization challenges of metabolic engineering, including high-dimensional and discontinuous solution spaces[32-34]. From pathway design to target prediction, phenotype simulation, and model parameter optimization, these algorithms provide effective tools for enhancing target product yields, optimizing strain performance, and uncovering metabolic network mechanisms. However, *in silico* strain design by combining Meta-heuristic algorithms and ecModels remains underexplored.

In this study, we present MetaStrain, a novel strain design framework that synergistically integrates ecModels with advanced meta-heuristic algorithms to identify non-intuitive combinatorial gene editing

strategies for maximizing product yields in metabolic engineering. This integration enables MetaStrain to achieve high computational efficiency with stable convergence, facilitating the rapid and accurate identification of optimal gene editing strategies involving over-expression, knockdown, and knockout. The framework demonstrates strong adaptability and generalization, being compatible with all ecModels developed using the GECKO framework, which broadens its application across diverse species and product targets. Furthermore, MetaStrain is designed for continuous upgradability, allowing seamless integration with state-of-the-art search algorithms and gene dimensionality reduction tools, ensuring its scalability and relevance for future advancements. Notably, it provides diverse and experimentally feasible gene editing strategies while effectively controlling the number of targets, balancing biological feasibility with cost-efficiency. By addressing the computational challenges of large-scale combinatorial optimization and offering comprehensive analyses of metabolic network phenotypes from a target editing perspective, MetaStrain represents a significant advancement in strain optimization, providing a powerful and innovative approach for reprogramming complex biochemical systems to achieve enhanced production metrics.

Results

Leveraging meta-heuristic algorithm on ecModels for non-intuitive combined target prediction

The MetaStrain comprises two primary modules (Figure 1a, b): the initial screening module, and the search module. In the first module, we used the first step of our previous algorithm ecFactory to reduce the dimensionality of enzyme-encoding genes, which will provide the potential gene list for next-step optimization.

In the searching module, each individual represents a set of gene editing strategies and ecMOMA (see Methods and Materials) is used to simulate the phenotype of experimentally modified strains (Figure 1c), with target product yield (Eq. 4) serving as the fitness value. A binary-coded GA (binGA) was firstly employed for rapid combinatorial target prediction, depending on the results of dimension reduction and annotation from the first module. Alternatively, a symbolic-coded GA (symGA) and JADE were leveraged to search the optimal combinations of gene manipulations based on the result of dimension reduction from the first module. While this approach demands more computational resources, it compensates for potential inaccuracies in the preceding dimension reduction module. This allows for a broader consideration of target interactions, overcoming limitations imposed by annotation results, and is well-suited for integration with dimension reduction algorithms that cannot annotate targets. During the search phase, an elite retention strategy can be incorporated to accelerate convergence. Additionally, target numbers are controlled through different fitness function designs or parameter adjustments (Methods and Materials).

To validate the stability and effectiveness of the MetaStrain framework, we systematically evaluated its performance in designing microbial cell factories for various target products. For each case, we conducted multiple independent optimization runs and selected the most promising gene editing strategy based on a comprehensive scoring (Eq. 7) of the predicted outcomes.

MetaStrain can efficiently identified combinatorial targets to enhance production using binGA

Identifying optimal combinations of genetic modifications is a formidable challenge in metabolic engineering, often plagued by combinatorial explosion. To assess the effectiveness of our algorithm in navigating this vast search space, we benchmarked MetaStrain's performance using 2-phenylethanol (2-PE), which is a high-value aromatic compound with wide applications in the food, cosmetics, and pharmaceutical industries[35], as a test case. Three-dimensional phase plane analysis revealed significant metabolic trade-offs between 2-PE production, biomass synthesis, and glucose consumption (Figure 2a), highlighting the need for rational adjustments to the metabolic network to optimize resource allocation without compromising cellular viability. Simulation using the ecYeast8.3.4 indicated a theoretical maximum flux of 4.9009 $\mathrm{mmol\ gDWh^{-1}}$ for 2-PE production, with an initial fitness value of 0.1231. Based on ecFSEOF[24], 52 candidate genes were identified (Table 2), and editing strategies for these genes were annotated. Single-target simulations were conducted based on dimension reduction and annotation information. However, the results showed that most single-gene edits had no effect, and some even reduced the yield of the target product (Figure 2b), highlighting the complex interaction mechanisms among genes. For instance, certain targets may be ineffective or detrimental when acting alone but exhibit a synergistic effect when combined with another target. These findings underscoring the necessity of exploring combinatorial target optimization strategies.

Using binGA, combinatorial target optimization was performed on the candidate genes. The algorithm demonstrated robust convergence results for both fitness functions, with and without a regularization term, successfully identifying multiple biologically meaningful editing strategies (Figure 2c, d). The binGA search using two fitness functions both employed elitism strategy. Among these, a strategy involving the knockdown of two targets—*CDC19* and *FUM1*—was repeatedly predicted (Figure 2e). In the metabolic pathway, *CDC19* encodes pyruvate kinase, a key enzyme in glycolysis that catalyzes the irreversible conversion of phosphoenolpyruvate to pyruvate. Knocking down *CDC19* reduces the glycolytic flux, thereby decreasing the input of pyruvate into the tricarboxylic acid (TCA) cycle. This increases the availability of the precursor phenylalanine, promoting 2-PE biosynthesis. On the other hand, *FUM1* encodes fumarase, which catalyzes the conversion of fumarate to malate in the TCA cycle. Knocking down *FUM1* weakens TCA cycle activity, reducing excessive competition in energy metabolism and

further optimizing carbon flux allocation toward the target product. The combined knockdown of *CDC19* and *FUM1* not only decreases resource consumption by competing metabolic pathways but also synergistically enhances carbon flux allocation to the target metabolic pathway. This strategy effectively reshapes global resource distribution in the metabolic network, providing a clear combinatorial target for efficient biosynthesis of 2-PE (Table 3). Additionally, the algorithm predicted strategies involving knockdown of targets such as *ARO8* and *ACO1*, offering further insights into non-intuitive targets for wet-lab validation.

MetaStrain showed better performance in a larger search space using JADE

In further optimization efforts, we explored a scenario where gene editing methods were not annotated using ecFSEOF, relying solely on the reduced-dimensional gene set for combinatorial target search. This strategy significantly expanded the search space while increasing the complexity of the search landscape, imposing higher demands on algorithm performance. Initially, we employed symGA for the search. Despite incorporating an elite retention strategy, symGA demonstrated fast convergence in the expanded search space (Supplemental Figure S1). Its stability across ten independent experiments was low (Figure 3a, Supplemental Table S1). Statistical analysis of the results revealed no single target consistently predicted across all ten runs, and many genes exhibited diverse editing strategies in different combinations (Figure 3b). This inconsistency shows the limitations of symGA in finding stable solutions within a complex search landscape.

Subsequently, JADE was introduced and exhibited remarkable performance advantages in large-scale search spaces (Figure 3c). Results from ten independent experiments showed that JADE achieved stable convergence and generated highly consistent optimization strategies. Statistical analysis revealed that seven genes were consistently identified as targets across all optimization results, with almost identical editing strategies. This high level of consistency supports the biological feasibility of the predicted strategies, providing a solid foundation for experimental design (Figure 3d). Additionally, JADE predicted diverse gene editing strategies. For example, three strategies included the overexpression of *ICL1*, which encodes isocitrate lyase, a key enzyme in the glyoxylate shunt pathway. Combined with control over other pathways, this optimized carbon flux distribution in *S. cerevisiae*, enhancing 2-PE synthesis. JADE's adaptive parameter control further improved optimization efficiency by eliminating the need for manual parameter tuning.

In the comparison of convergence performance across algorithms, the convergence curves of binGA, symGA, symGA without an elite retention strategy, and JADE were plotted on the same graph (Figure 4a), using $F1$ (Eq. 5) as the fitness function. Results showed that binGA, with its simple search strategy, converged quickly, achieving good results within fewer iterations. In contrast, JADE exhibited higher

convergence quality due to its powerful global search capabilities, optimizing the objective function further in complex search landscapes. Details of the strategies with highest fitness value were presented in Table 3.

MetaStrain can effectively control the number of combined targets and capture validated metabolic engineering targets in the prediction result

Controlling the number of gene targets is critical for metabolic engineering, as excessive targets can increase experimental costs and reduce strain modification feasibility. This study analyzed the performance of different algorithms to verify MetaStrain's effectiveness in controlling the number of predicted targets. To evaluate the control of target numbers, the number of targets from ten independent experiments using binGA (with and without regularized fitness functions) and JADE, was statistically analyzed (Figure 4b). Results indicated that binGA combined with a regularization strategy effectively limited the number of predicted targets. In contrast, JADE, relying on its strong search performance, automatically excluded redundant targets during iterations, yielding strategies that were acceptable for wet-lab validation. Because symGA's convergence is unstable, it was not included in the comparison. However, fewer targets do not necessarily lead to better strategies. Enforcing strict limits on target numbers may hinder the algorithm's ability to explore potentially superior solutions, which could negatively impact rational MCF design. From a biological perspective, MetaStrain's optimized results demonstrate significant advantages. By effectively controlling the number of targets, the proposed strategies reduce the complexity of wet-lab experiments while enhancing the feasibility and cost-efficiency of strain modifications.

The effectiveness of metabolic engineering optimization algorithms hinges on their ability to identify experimentally validated gene targets while proposing innovative and insightful strategies. In this study, a comparative analysis of the predictions made by binGA (using regularized fitness functions) and JADE demonstrated that both algorithms frequently identified key targets. JADE further exhibited superior predictive performance, with its high-frequency targets encompassing those identified by binGA (Figure 4c). Moreover, the overlapping gene editing strategies were entirely consistent, such as the knockdown of *CDC19* and *FUM1*. Specifically, *CDC19*, an experimentally validated target, was consistently selected and annotated as a knockdown target in both binGA and JADE results. Additionally, the dimension reduction performed by ecFSEOF ensured that the initial candidate pool of 52 genes included seven experimentally verified targets, with the proposed combinatorial target strategies containing1-3 validated targets, enhancing the biological reliability of the predictions. Beyond the validated targets, both JADE and binGA also proposed numerous novel gene targets that have not been reported in the current literature[24, 36]. These targets exhibit significant engineering value due to their potential roles in redistributing carbon flux or reducing network redundancy to enhance target product yield. Such

non-intuitive predictions highlight the power of meta-heuristic algorithms in optimizing complex networks, uncovering latent targets that conventional empirical methods fail to identify. MetaStrain offers a new perspective for understanding the synergistic interactions within metabolic networks, advancing the development of efficient MCF design.

MetaStrain performs well in generalization experiments

To assess the generalization ability of our approach, we selected metabolic engineering studies of multiple products to validate the predictive power of the algorithm. Spermidine is a biologically significant polyamine compound with potential in anti-aging and cell growth regulation[37]. Phase plane analysis revealed significant competition between spermidine production and biomass formation, emphasizing that optimizing the allocation of sugar metabolism resources is critical for improving spermidine yield. The ecFSEOF analysis identified 86 candidate genes (Table 2), but single-target simulations showed no significant yield improvement, highlighting the necessity for combinatorial targets (Supplemental Figure S2a, b).

We employed JADE to perform combinatorial target optimization. The searching results demonstrated stable convergence (Supplemental Figure S2c). Statistical analysis of ten independent experiments revealed that two targets were consistently selected in all predictions with identical editing directions: the knockout of *EMI2* and the overexpression of *SPE3* (Supplemental Figure S2d). *SPE3*, an experimentally validated target, encodes arginine decarboxylase, which catalyzes the conversion of arginine to ornithine—a crucial upstream step in spermidine biosynthesis[38]. Overexpression of *SPE3* significantly enhanced the supply of precursors for spermidine synthesis. Conversely, the knockout of *EMI2* reduced competitive effects in ornithine catabolism, redirecting more resources toward the spermidine synthesis pathway and further optimizing carbon and nitrogen flux allocation. Additionally, JADE predicted other potential targets, such as *APT1* and *CAR2*, both of which are experimentally validated gene targets. Overexpression of *APT1* enhances ATP synthesis, indirectly supporting energy-intensive biosynthetic pathways, while *CAR2* regulates nitrogen metabolism balance in the ornithine cycle, providing a more stable environment for spermidine production. The results demonstrate that MetaStrain captures not only experimentally validated targets but also non-intuitive combinatorial targets related to the metabolism of the target product (Table 4). This robust generalization ability underscores the algorithm's applicability across diverse target product optimizations, offering a versatile solution for the complex optimization problems commonly encountered in metabolic networks.

Discussion

The MetaStrain algorithm proposed in this study combines ecModels with efficient meta-heuristic algorithms, offering an innovative solution to the combinatorial target optimization challenges in complex

metabolic engineering. Its exceptional performance across multiple optimization tasks demonstrates the algorithm's ability to identify non-intuitive and innovative gene editing strategies while ensuring result stability, thus providing critical support for developing efficient microbial cell factories. While GA provide simplicity and efficiency, particularly in smaller search spaces, the JADE significantly enhances global search capability through adaptive parameter control and external archive mechanisms, achieving efficient convergence even in large and complex search spaces. Notably, JADE accurately identifies experimentally validated targets and proposes novel targets, underscoring MetaStrain’s potential in integrating existing knowledge with uncovering new strategies. The strength of MetaStrain lies in its ability to transition from single-target to combinatorial-target optimization, providing a comprehensive means of addressing complex metabolic network competition and reshaping carbon flux distribution. Its successful adaptation to the replacement of the target product further reinforces its broad applicability.

Despite MetaStrain's promising performance, there remains room for improvement. Incorporating multi-objective optimization algorithms could enable simultaneous optimization of multiple indicators, enhancing practical applicability. Additionally, integrating surrogate models, such as machine learning-based metabolic flux prediction models, could accelerate the computational process, particularly in large-scale search spaces. Exploring the integration of more meta-heuristic algorithms and dynamic environment optimization mechanisms may further enhance the algorithm's adaptability and robustness.

In conclusion, the MetaStrain algorithm offers a novel chance for optimizing complex metabolic networks, with its innovation, stability, and adaptability, laying a solid foundation for advancing metabolic engineering. Combining MetaStrain with high-throughput experimental platforms could extend its impact from computational optimization to experimental validation, creating a closed-loop design process that tightly couples computation and experimentation. This synergistic approach will not only accelerate the development of microbial cell factories but also drive the widespread application of metabolic engineering in sustainable biomanufacturing solutions.

Materials and Methods

Framework of MetaStrain

The MetaStrain workflow (Figure 1) comprises four key steps:

1) Model preprocessing and reference flux acquisition

For different microorganisms, ecModels are first adjusted and validated to accurately simulate metabolic states under different environmental conditions. This includes enabling oxygen, major carbon/nitrogen sources (with glucose as the sole carbon source in this study) and setting the upper bounds of reactions that initially defined as infinite to 1000. To enhance the resolution of protein synthesis-related fluxes,

model precision is set to 10^{-9}. Essential genes and redundant metabolic pathways are removed. For exogenous products, relevant pathways are incorporated into the model, including adding new reactions, updating existing reactions, implementing gene manipulations, and setting constraints. The wild-type strain phenotype is then obtained through FBA by enforcing a minimum production level of the target product.

2) Gene dimension reduction and annotation

The ecFSEOF[24] method is used to perform gene dimension reduction and annotation for genes encoding enzymes in the model. To ensure that the randomly initialized individuals in a relatively small population can sufficiently cover the defined decision space, the number of candidate genes is reduced to fewer than 100. This significantly reduces the complexity of the search space, providing an efficient starting point for combinatorial target optimization.

3) Optimization process and algorithm implementation

MetaStrain integrates an efficient meta-heuristic optimization framework, using the yield of the target product as the objective function and encoding individuals according to the requirements of different algorithms. The main steps include population initialization fitness evaluation, and evolutionary operations. Additional techniques, such as elite retention to accelerate convergence, regularization terms in the fitness function to control the number of targets, adaptive parameter control, and external archives, can further enhance algorithm performance.

4) Result Analysis and Validation

Due to the stochastic nature of the algorithm, ten independent experiments are conducted for each target product in a specific strain. The optimization results are scored using (Eq. 7) and statistically analyzed to identify combinatorial targets with strong stability and biological significance. The results are further validated against experimentally verified targets comparation and pathway mechanism studies to assess their rationality and feasibility.

ecModel of *S. cerevisiae*

GEMs are widely used to represent the metabolic networks of microorganisms by integrating genes, enzymes, and reactions through gene-protein-reaction (GPR) rules. These models encode microbial metabolism in a $M \times N$ dimensional stoichiometric matrix ($\mathbf{S}$ matrix), where rows represent metabolites, columns correspond to reactions, and matrix elements denote stoichiometric coefficients. The GECKO method expands GEMs by introducing additional rows and columns into the $\mathbf{S}$ matrix to explicitly include enzymes as reaction components. This approach highlights a fundamental constraint on metabolic fluxes (1): each flux (v_j) cannot exceed the maximum enzymatic rate ($v_{\max}$), which is determined by the intracellular concentration of the enzyme and its turnover number.

$$v_j \le k_{cat}^{ij} \cdot [E_i], \tag{1}$$

Here, $[E_i]$ denotes the intracellular concentration of each enzyme E_i, and k_{cat}^{ij} is the turnover rate. This quantitative method imposes soft constraints at the enzyme level, integrating proteomics data. In cases where detailed proteomics data are unavailable, a global enzyme mass constraint, similar to the FBAwMC[39] method, is applied. This introduces a pseudo-metabolite termed the "protein pool", to ensure that all metabolic reaction fluxes remain within biologically reasonable limits. For this study, we used the ecModel of *S. cerevisiae* ecYeast8.3.4, a well-characterized industrial chassis for metabolic engineering. Developed and refined by Lu et al.[40] (Table 1).

Phenotype simulation of wild-type and mutant strains

Flux balance analysis

FBA employs linear programming techniques, such as the simplex method, to solve optimization problems. FBA predicts the steady-state metabolic phenotype of a microorganism under specified environmental conditions, without relying on complex kinetic parameters. For single-celled organisms, this objective often aligns with maximizing biomass synthesis, under the assumption that these organisms have evolved to achieve optimal growth performance.

$$\max\ v_{\text{biomass}} \quad s.t. \begin{cases} \mathbf{Sv} = 0 \\ v_j^{\min} \le v_j \le v_j^{\max} \\ \forall i \in M,\ \forall j \in N \end{cases}, \tag{2}$$

Here, v_{biomass} represents the flux of biomass production. $\mathbf{Sv} = 0$ provides a global equality constraint for the model, ensuring mass conservation. $v_j^{\min}$ and $v_j^{\max}$ represent the minimum and maximum bounds of a reaction v_j. The problem can be solved using a linear programming solver, such as Gurobi[41].

In this study, the wild-type strain's metabolic phenotype serves as the baseline for subsequent computational analyses, which was predicted by a single-step FBA-based approach. First, the flux range of product was first set approximately half of its theoretical maximum flux, ensuring that the simulated strain maintains a certain production level. Biomass synthesis was then maximized as the objective function for solving.

Phase plane analysis

Phase plane analysis can be performed by gradually varying the target flux and constraint conditions. It provides a comprehensive understanding of the global characteristics of metabolic networks, particularly trade-off relationships between two or more metabolic fluxes. In this study, the StrainDesign[42] toolbox was used to construct phase planes for a target product, biomass synthesis, and glucose consumption. This

method allows for quick and intuitive insights into nonlinear couplings between fluxes of interest, identifying the factors influencing product yields.

Minimization of Metabolic Adjustment in ecModel

Genetically engineered strains often exhibit suboptimal metabolic flux distributions due to the absence of natural evolutionary processes, with an initial tendency to resemble the wild-type flux. To predict these suboptimal states, the MOMA method minimizes the Euclidean distance (Eq. 3) between the flux profiles of wild-type and mutant strains, formulated as a quadratic programming problem. While effective for standard GEMs, MOMA requires adaptation for ecModels due to their unique structure. Here, we introduce an improved method, ecMOMA.

$$D(\mathbf{w},\mathbf{v}) = \sqrt{\sum_{i=1}^{N}\left(w_i - v_i\right)^2}, \tag{3}$$

Here, $\mathbf{w}$ represents the metabolic flux vector of the wild-type strain, serving as the reference vector in the algorithm, and $\mathbf{v}$ represents the flux distribution of the mutant strain.

In ecModels, each enzyme is associated with a “synthesis reaction” and an “arm reaction” to ensure total flux constraints. However, the flux magnitudes of enzyme-synthesis reactions significantly lower than those of the original metabolic reactions. This discrepancy causes MOMA results to be dominated by the original metabolic reactions, and the introduced arm reactions become redundant, leading to double-counting of certain metabolic fluxes. To resolve this, ecMOMA refines the distance function by excluding enzyme-synthesis reactions, arm reactions, and protein pool exchange reactions from the flux space. This adaptation improves simulation accuracy, reduces computational redundancy, and enhances efficiency in both time and memory usage.

Gene dimension reduction and labeling using ecFSEOF

The ecFSEOF algorithm was employed to identify gene editing targets. This method integrates the features of GECKO, extends its application to ecModels, and incorporates a scoring function to prioritize genes for overexpression, knockdown, or knockout.

In this study, the algorithm sets the biomass synthesis flux to its theoretical maximum and progressively decreases it across predefined levels to explore metabolic changes. Flux values are obtained through FBA with the target product yield as the objective function, followed by a two-step scoring process to identify potential engineering targets. For a detailed description of the scoring methods and criteria for gene prioritization, please refer to the supplementary materials (Supplemental Material).

Characterization of Gene Editing in Models

In MCF design, representing gene editing operations—specifically, overexpression, knockdown, and knockout of genes, and determining the extent of these edits—remains challenging. The ecModels address this by modulating enzyme usage, as they treat enzymes as key components directly involved in metabolic reactions. Enzyme synthesis reactions establish a one-to-one correspondence between enzymes and genes, while other reactions represent enzyme utilization. These relationships are embedded in the metabolic network and adhere to mass conservation principles.

In ecYeast, gene knockout was implemented by setting the flux bounds of the corresponding enzyme synthesis reaction to zero, effectively deactivating the pathway. For overexpression and knockdown, flux bounds were manually adjusted based on reference enzyme usage values derived from the wild-type phenotype. Overexpression was defined as increasing the lower flux bound to four times its reference value, while knockdown reduced the upper flux bound to half of its reference value. These adjustments were informed by literature and practical considerations to ensure rational and effective interventions[17]. To address the issue of zero flux predictions for many reactions due to model precision limitations, a minimum flux value ($\varepsilon = 4\times10^{-9}$) was introduced. For genes targeted for overexpression with an initial flux of zero, the lower flux bound was set to 4ε, ensuring feasibility within the model's precision constraint.

Meta-Heuristic Algorithm

Meta-heuristic algorithms operate through iterative processes and parallel searches to identify satisfactory solutions, rather than absolute optimal. These algorithms are especially suited for global optimization problems, as they balance exploration and exploitation. While GA and DE share similar iterative processes, they differ in their evolutionary operators and auxiliary search strategies. In this study, GA and adaptive differential evolution with optional external archives (JADE)[43] were employed for combinatorial target searches. GA was divided into two forms, binary-coded (binGA) and symbolic-coded (symGA), based on whether gene annotations were available. While JADE's adaptive parameter control and external archives increase computational overhead, they offer superior global optimization capabilities and can achieve effective results without relying on gene annotations.

Innovatively, we designed parameters within GA and JADE initializations to ensure population uniformity and individual feasibility. Appropriate thresholds maintained a low number of targets, ensuring experimental economic viability. Individual encoding complied with algorithm requirements; for example, JADE operated within a continuous space. These initialization strategies avoid excessive interventions, thus promoting feasible and economically viable solutions. In binGA and symGA, a roulette wheel

selection strategy was used to select candidate individuals, and an elite retention strategy ensured the direct inheritance of top performers to the next generation. Threshold control during single-point crossover and mutation further optimized population updates (see Supplemental Material). JADE introduces unique mutation strategies DE/current-to-*p*best*s* , optional external archive and adaptive parameter control to improve its global optimization capabilities. The external archive adds diversity during iterations, ensuring effective search direction. Detailed JADE operating procedures and equations are available in the supplemental material.

Evaluation of Individuals

In meta-heuristic algorithms, the quality of each individual is assessed through its fitness, with higher fitness values indicating better solutions. TRY are valuable economic indicators in fermentation processes. In this study, the yield of the target product is selected as the objective function, defined as:

$$Yeild = \frac{flux_{\text{product}}}{flux_{\text{glucose}}} \tag{4}$$

During the evaluation process, the wild-type strain phenotype is first obtained as a reference flux through FBA of the original model. Then, the ecModel is modified according to the gene editing strategy represented by the individual. These two inputs are used in ecMOMA to compute the mutant strain phenotype and subsequently calculate the target product yield. Notably, a minimum growth capacity is ensured for mutant strains, with biomass synthesis flux constrained to no less than 0.1 mmol gDWh^{-1} .

We desinged two types of fitness function. First, we directly took yield as the fitness value (Eq. 5). Then we consider adding a regularization term to the function to control the number of combined targets (Eq. 6).

$$F_1 = Yeild \ , \tag{5}$$

$$F_2 = Yeild + \frac{1}{10000} N_0 , \tag{6}$$

Here, N_0 denotes the number of dimensions encoded as 0 in the individual, and is scaled to adjust the influence of the target count on the fitness value. Other regularization terms, like the double hook function, show the same result as the linear term while requiring additional computing resources.

It is important to note that the ecMOMA calculation is essentially a quadratic programming problem, and models with extensive gene edits may not always yield optimal solutions. The solver may return statuses such as “optimal”, “infeasible” or “unbounded”. To ensure the algorithm proceeds smoothly, non-optimal cases are assigned a small nonzero fitness value (0.001). This assignment ensures that these individuals have a low probability of selection without significantly affecting population quality, thereby increasing the algorithm’s diversity.

Ranking of Optimal Individuals

Through repeated experiments, we can obtain the optimal editing strategies found in each search, along with their corresponding fitness values. Following the search, strategies were ranked based on three combined criteria: the regularized fitness values, the total number of targets, and the number of experimentally validated targets. By applying weighted summation to these components, we derived a prioritization strategy to guide subsequent wet lab experiments.

$$R_{\text{score}} = Yield + \frac{1}{10000} N_0 + \frac{1}{1000} N_{\text{exp}}, \tag{7}$$

where N_{exp} represents the number of experimentally validated targets in the strategy. This comprehensive ranking approach ensures that the selected gene editing strategies not only maximize yield but also align with experimental validation results.

Key points

- This study introduces the MetaStrain algorithm, a novel optimization framework that integrates ecModels with efficient meta-heuristic algorithms to identify non-intuitive combinatorial gene targets, significantly enhancing product yields in metabolic engineering.
- MetaStrain demonstrates high computational efficiency and stable convergence, effectively balancing biological feasibility with cost-efficiency while providing diverse and experimentally feasible gene editing strategies for strain optimization.
- The framework's adaptability allows it to generalize across various metabolic targets, as validated through significant improvements in 2-phenylethanol and spermidine production in Saccharomyces cerevisiae, showcasing its robustness in complex metabolic networks.
- By addressing the combinatorial challenges inherent in metabolic engineering, MetaStrain offers a powerful computational tool that advances the design of high-performance microbial cell factories, paving the way for sustainable biomanufacturing solutions.

Acknowledgement

This work was financially supported by the National key research and development program of China (2020YFA0908300), Shanghai Municipal Science and Technology Major Project, and grant 22208211 and 22378263 from the National Natural Science Foundation of China (NSFC).

Data and code availability

The MetaStrain method is available as a standalone tool at the public repository: https://github.com/MBLiao/MetaStrain.

Conflict of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Jullesson D, David F, Pfleger B et al. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals, Biotechnology Advances 2015;33:1395-1402.
2. Zhou YJ, Buijs NA, Zhu Z et al. Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories, Nat Commun 2016;7:11709.
3. Nielsen J, Larsson C, van Maris A et al. Metabolic engineering of yeast for production of fuels and chemicals, Curr Opin Biotechnol 2013;24:398-404.
4. Ro DK, Paradise EM, Ouellet M et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast, Nature 2006;440:940-943.
5. Wei Y, Bergenholm D, Gossing M et al. Expression of cocoa genes in Saccharomyces cerevisiae improves cocoa butter production, Microb Cell Fact 2018;17:11.
6. Duman-Özdamar ZE, Binay B. Production of Industrial Enzymes via Pichia pastoris as a Cell Factory in Bioreactor: Current Status and Future Aspects, Protein J 2021;40:367-376.
7. Nielsen J, Keasling JD. Engineering Cellular Metabolism, Cell 2016;164:1185-1197.
8. Volk MJ, Tran VG, Tan S-, I et al. Metabolic Engineering: Methodologies and Applications, Chemical Reviews 2023;123:5521-5570.
9. Konzock O, Nielsen J. TRYing to evaluate production costs in microbial biotechnology, Trends Biotechnol 2024;42:1339-1347.
10. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods, Nature Reviews Microbiology 2012;10:291-305.
11. O'Brien EJ, Monk JM, Palsson BO. Using Genome-scale Models to Predict Biological Capabilities, Cell 2015;161:971-987.
12. Sanchez BJ, Zhang C, Nilsson A et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints, Molecular Systems Biology 2017;13.
13. Basan M, Hui S, Okano H et al. Overflow metabolism in Escherichia coli results from efficient proteome allocation, Nature 2015;528:99-104.
14. Nilsson A, Nielsen J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase, Scientific Reports 2016;6.
15. Shlomi T, Benyamini T, Gottlieb E et al. Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect, Plos Computational Biology 2011;7.
16. Orth JD, Thiele I, Palsson BO. What is flux balance analysis?, Nature Biotechnology 2010;28:245-248.
17. Segrè D, Vitkup D, Church G. Analysis of optimality in natural and perturbed metabolic networks, Proceedings of the National Academy of Sciences of the United States of America 2002;99:15112-15117.
18. Burgard A, Pharkya P, Maranas C. OptKnock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, Biotechnology and bioengineering 2003;84:647-657.

19. Patil KR, Rocha I, Forster J et al. Evolutionary programming as a platform for in silico metabolic engineering, Bmc Bioinformatics 2005;6:308.

20. Choi HS, Lee SY, Kim TY et al. In silico identification of gene amplification targets for improvement of lycopene production, Appl Environ Microbiol 2010;76:3097-3105.

21. Jensen K, Broeken V, Hansen ASL et al. OptCouple: Joint simulation of gene knockouts, insertions and medium modifications for prediction of growth-coupled strain designs, Metabolic engineering communications 2019;8:e00087-e00087.

22. Ranganathan S, Suthers PF, Maranas CD. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions, PLoS Comput Biol 2010;6:e1000744.

23. Jiang S, Otero-Muras I, Banga JR et al. OptDesign: Identifying Optimum Design Strategies in Strain Engineering for Biochemical Production, Acs Synthetic Biology 2022;11:1531-1541.

24. Domenzain I, Lu Y, Wang H et al. Computational biology predicts metabolic engineering targets for increased production of 103 valuable chemicals in yeast, Proc Natl Acad Sci U S A 2025;122:e2417322122.

25. Sabzevari M, Szedmak S, Penttila M et al. Strain design optimization using reinforcement learning, Plos Computational Biology 2022;18.

26. Zhang J, Petersen SD, Radivojevic T et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism, Nature Communications 2020;11.

27. Lu J, Bi X, Liu Y et al. In silico cell factory design driven by comprehensive genome-scale metabolic models: development and challenges, Systems Microbiology and Biomanufacturing 2023;3:207-222.

28. Holland JH. Genetic Algorithms, Scientific American 1992;267:66-72.

29. Das S, Suganthan PN. Differential Evolution: A Survey of the State-of-the-Art, Ieee Transactions on Evolutionary Computation 2011;15:4-31.

30. Huang C, Li Y, Yao X. A Survey of Automatic Parameter Tuning Methods for Metaheuristics, Ieee Transactions on Evolutionary Computation 2020;24:201-216.

31. Deb K, Pratap A, Agarwal S et al. A fast and elitist multiobjective genetic algorithm: NSGA-II, Ieee Transactions on Evolutionary Computation 2002;6:182-197.

32. Wang X, Wang Y, Tang L et al. Multiobjective Ensemble Learning With Multiscale Data for Product Quality Prediction in Iron and Steel Industry, Ieee Transactions on Evolutionary Computation 2024;28:1099-1113.

33. Liu W, Zhang Y, Liu K et al. Evolutionary Multiobjective Optimization for Large-Scale Portfolio Selection With Both Random and Uncertain Returns, Ieee Transactions on Evolutionary Computation 2025;29:76-90.

34. Yu Y, Tang Q, Jiang Q et al. A Deep Reinforcement Learning-Assisted Multimodal Multi-Objective Bi-Level Optimization Method for Multi-Robot Task Allocation, Ieee Transactions on Evolutionary Computation 2025:1-1.

35. Wang Y, Zhang H, Lu X et al. Advances in 2-phenylethanol production from engineered microorganisms, Biotechnol Adv 2019;37:403-409.

36. Yan Wei GAOHJYQXZJDWZWXINFJM. Research progress in 2-phenylethanol production through biological processes, Synthetic Biology Journal 2022;2:1030-1045.

37. Madeo F, Eisenberg T, Pietrocola F et al. Spermidine in health and disease, Science 2018;359.

38. Friesen H, Tanny JC, Segall J. Spe3, which encodes spermidine synthase, is required for full repression through NRE(DIT) in Saccharomyces cerevisiae, Genetics 1998;150:59-73.

39. Beg QK, Vazquez A, Ernst J et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity, Proc Natl Acad Sci U S A 2007;104:12663-12668.

40. Lu H, Li F, Sanchez BJ et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism, Nat Commun 2019;10:3586.

41. Gurobi Optimization LLC. Gurobi Optimizer Reference Manual.

42. Schneider P, Bekiaris PS, von Kamp A et al. StrainDesign: a comprehensive Python package for computational design of metabolic networks, Bioinformatics 2022;38:4981-4983.

43. Zhang J, Sanderson AC. JADE: Adaptive Differential Evolution With Optional External Archive, Ieee Transactions on Evolutionary Computation 2009;13:945-958.

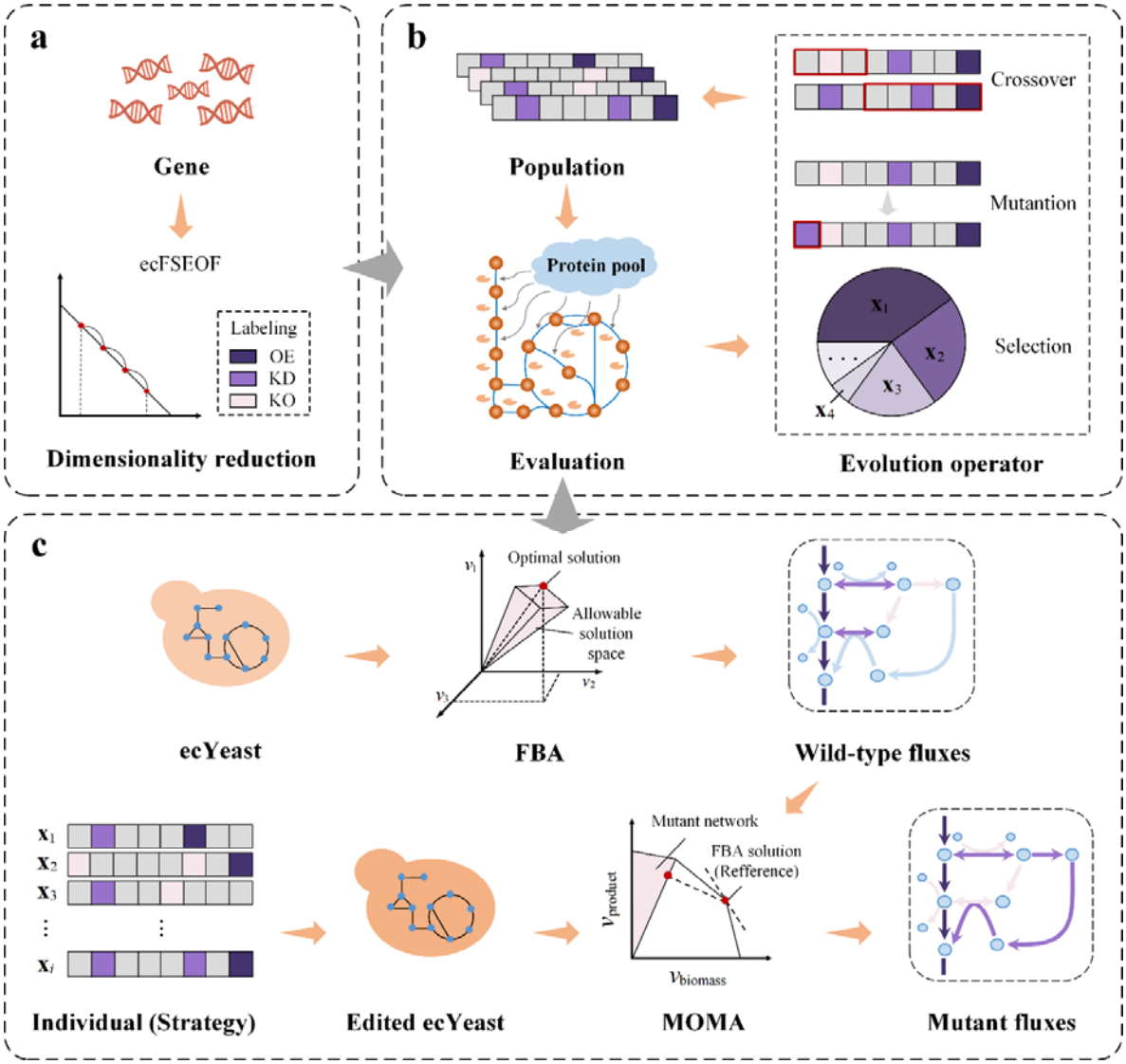

Figure 1 The computational framework of MetaStrain. The algorithm consists of two primary modules. (a) The initial screening module performs dimensionality reduction and annotation. In this stage, all genes encoding enzymes in the metabolic model are systematically analyzed to pre-screen potential targets. (b, c) The search module, which executes the core optimization task. (b) The general iterative workflow of a meta-heuristic algorithm, where a population of candidate designs undergoes mutation, crossover, and selection. (c) The details of the fitness evaluation for each individual design. The process involves simulating a reference flux state and then solving for the flux distribution of the genetically modified strain, ultimately yielding a fitness score for the iterative search.

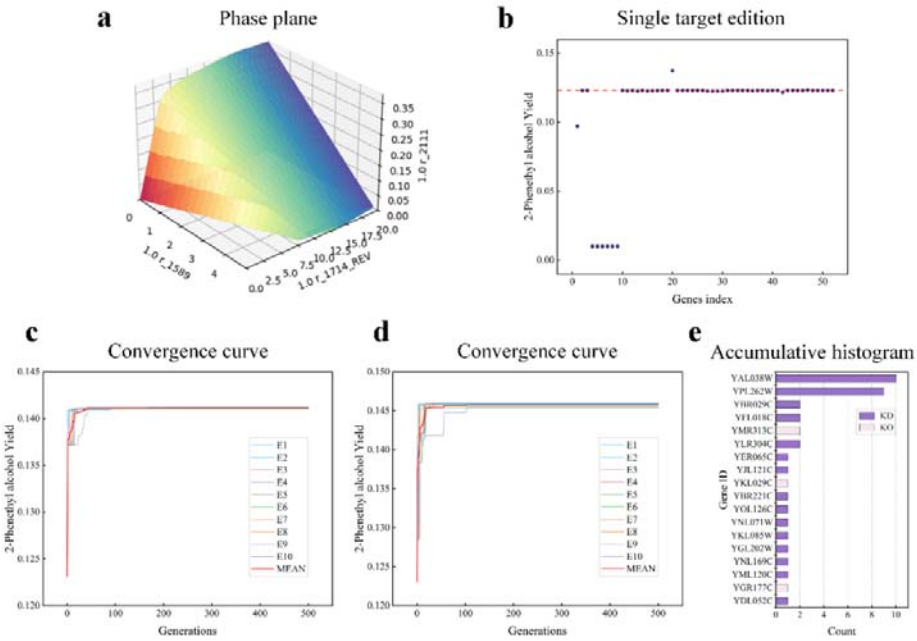

Figure 2 Identification of combinatorial targets for 2-PE production using binGA. (a) Three-dimensional phase plane analysis of the wild-type *S. cerevisiae* ecModel, illustrating the trade-offs between 2-PE production, biomass synthesis, and glucose uptake rate. The plot reveals a strong conflict between growth and product formation. (b) Predicted 2-PE production yields resulting from the overexpression, down-regulation, or knockout of single gene targets. (c, d) Convergence performance of the binGA search process. The plots compare the evolution of the population's average fitness over generations (c) with and (d) without the inclusion of a regularization term penalizing the number of gene modifications. (e) An accumulative histogram summarizing the frequency of each gene's appearance in the optimal solutions identified across ten independent binGA runs.

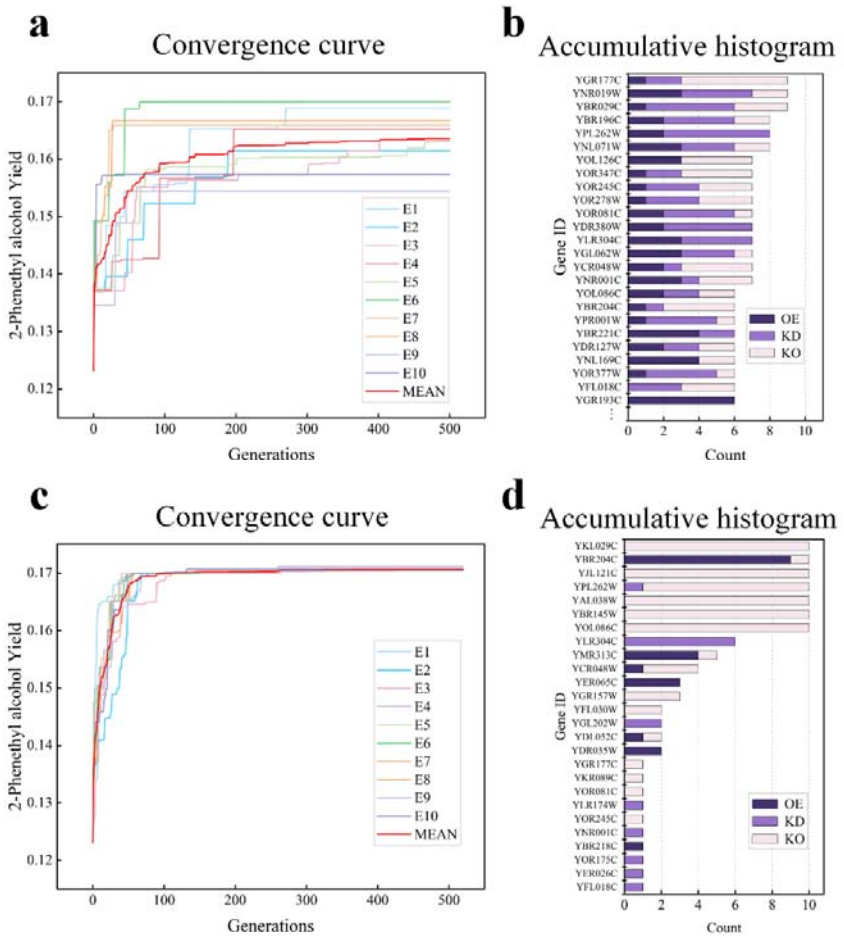

Figure 3 Performance comparison of symGA and JADE for 2-PE target identification in an expanded search space. This analysis was conducted using only the dimensionality reduction results, without the annotation of specific gene manipulation strategies, thereby significantly increasing the complexity of the optimization task. (a, b) Performance of the symGA algorithm, showing (a) the convergence curve of the optimization process and (b) an accumulative histogram of the optimal gene targets identified across ten independent runs. (c, d) Corresponding performance of the JADE algorithm, showing (c) its convergence profile and (d) the statistical distribution of its identified optimal targets.

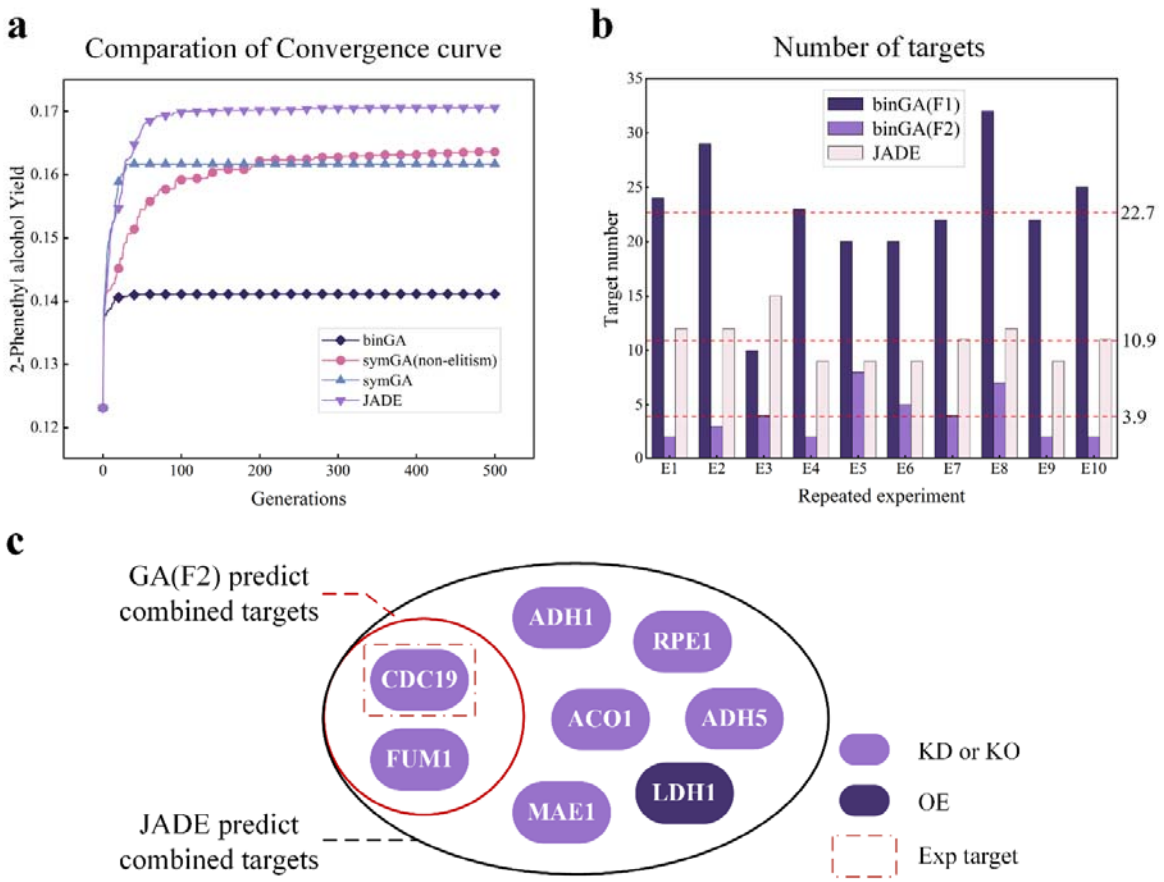

Figure 4 Comprehensive performance comparison and target analysis of different optimization algorithms. (a) Convergence profiles of the four algorithms evaluated in this study, illustrating their relative efficiency and effectiveness in reaching optimal solutions. (b) A bar chart comparing the number of genetic modifications required by the optimal strategies across ten independent runs. This highlights the ability of some algorithms to find more experimentally feasible solutions. (c) A Venn diagram illustrating the overlap between the predicted gene targets and a set of experimentally validated targets from literature.

Table 1 Details of the ecYeast 8.3.4

Content	Number
All reaction	8144
Metabolite	4180
Gene	1148
Gene that encode enzyme	968
Enzyme synthesis reaction	968
Enzyme with abundance data	968
Arm reaction	520
Metabolic reaction	6655
Protein pool	1
cell compartment	14

Table 2 The result of ecFSEOF

Operation	2-PE targets count	Spermidine targets count
OE	11	68
KD	35	0
KO	6	18
Total	52	86

Table 3 Details of the top-ranked gene editing strategies predicted by binGA (using regularized fitness function)

Algorithm	Individual (Strategy)	Fitness score	OE	KD	KO	Number of targets	Number of exp targets
binGA(F2)	$\mathbf{x}_1$	0.1470	\	**YAL038W**, YPL262W	\	2	1
JADE	$\mathbf{x}_2$	0.1778	**YDR035W**, YER065C, YBR204C	**YGL202W**, YLR304C	YOL086C, YBR145W, **YAL038W**, YPL262W, YJL121C, YKL029C	11	3
	$\mathbf{x}_3$	0.1768	YBR218C, YDL052C, YCR048W	YBR145W, YFL018C, **YGL202W**, YLR304C, YFL030W, YOR245C	YOL086C, **YAL038W**, YPL262W, YJL121C, YBR204C, YKL029C	15	2
	$\mathbf{x}_4$	0.1765	**YDR035W**, YER065C, YBR204C, YMR313C	YPL262W, YLR174W	YOL086C, YBR145W, **YAL038W**, YCR048W, YJL121C, YKL029C	12	2

Table 4 The ranked strategies obtained by JADE

Product	Individual (Strategy)	F1	F2	R_{score}	Number of targets	Number of exp targets
2-PE	$\mathbf{x}_1$	0.1707	0.1748	0.1778	11	3
	$\mathbf{x}_2$	0.1711	0.1748	0.1768	15	2
	$\mathbf{x}_3$	0.1705	0.1745	0.1765	12	2
	$\mathbf{x}_4$	0.1707	0.1750	0.1760	9	1
	$\mathbf{x}_5$	0.1705	0.1748	0.1758	9	1
	$\mathbf{x}_6$	0.1705	0.1748	0.1758	9	1
	$\mathbf{x}_7$	0.1705	0.1748	0.1758	9	1
	$\mathbf{x}_8$	0.1706	0.1746	0.1756	12	1
	$\mathbf{x}_9$	0.1705	0.1746	0.1756	11	1
	$\mathbf{x}_{10}$	0.1705	0.1745	0.1755	12	1
Spermidine	$\mathbf{x}_1$	0.1934	0.2001	0.2021	19	2
	$\mathbf{x}_2$	0.1918	0.1990	0.2010	14	2
	$\mathbf{x}_3$	0.1918	0.1989	0.2009	15	2
	$\mathbf{x}_4$	0.1919	0.1989	0.2009	16	2
	$\mathbf{x}_5$	0.1905	0.1974	0.1994	17	2
	$\mathbf{x}_6$	0.1910	0.1982	0.1992	14	1
	$\mathbf{x}_7$	0.1906	0.1972	0.1982	20	1
	$\mathbf{x}_8$	0.1853	0.1920	0.1950	19	3
	$\mathbf{x}_9$	0.1829	0.1899	0.1909	16	1
	$\mathbf{x}_{10}$	0.1806	0.1876	0.1886	16	1