ACS Synthetic Biology

pubs.acs.org/synthbio
Research Article

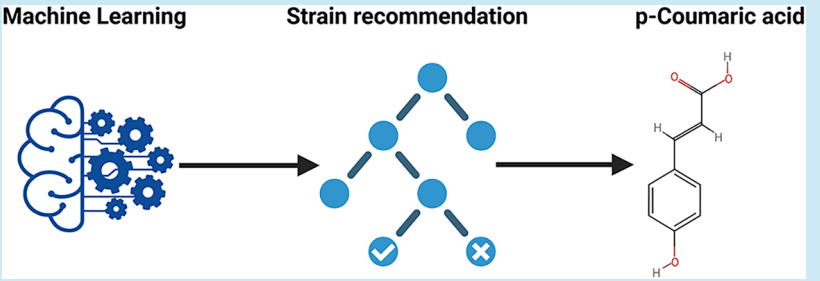
Machine-Learning-Assisted Pathway Optimization in Large Combinatorial Design Spaces: A *p*-Coumaric Acid Case Study

Published as part of ACS Synthetic Biology special issue "Artificial Intelligence for Synthetic Biology".

Paul van Lent, Rianne van der Hoek, Sara Moreno Paz, Irsan Kooi, Moniek Jonkers, Priscilla Zwartjens, Joep Schmitz, and Thomas Abeel*

Cite This: *ACS Synth. Biol.* 2026, 15, 3146–3157
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Combinatorial pathway optimization is a powerful approach in metabolic engineering to improve strain performance. While machine learning (ML) has shown promise in guiding the Design−Build−Test−Learn (DBTL) cycle, most applications have been limited to small design spaces, thereby restricting the potential of predictive and exploration-exploitation strategies. In this work, we applied two DBTL cycles to optimize *p*-coumaric acid production in *Saccharomyces cerevisiae*. The first cycle involved constructing a large combinatorial library of 18 genes and 20 promoters (170 million possible designs). In the second cycle, we employed a gradient bandit-based machine learning recommendation strategy, tuned to balance exploration and exploitation. Our results show that this balanced strategy outperforms greedy, feature importance-based approaches, leading to greater diversity in strain performance and improved top-producer identification. Notably, applying the same strategy to an alternative parent strain yielded the highest *p*-coumaric acid titer (1.23 g/L), a 2.37-fold improvement over the original. These findings highlight the value of ML-guided exploration in large design spaces and demonstrate that balancing exploration and exploitation is critical for successful strain optimization.

KEYWORDS: *metabolic engineering, combinatorial pathway optimization, strain recommendation, machine learning*

INTRODUCTION

Transitioning chemical production to sustainable processes is critical for addressing climate and biodiversity challenges.[7] Microbial cell factories offer an environmentally friendly alternative for producing pharmaceuticals, feed additives, and bulk chemicals.[18] However, the lengthy and costly development of proof-of-concept strains limits their widespread adoption.[10,22]

Advances in genome engineering and high-throughput screening have paved the way for combinatorial pathway optimization.[14] This strategy targets multiple pathway elements simultaneously to identify configurations that maximize titer, rate, and yield (TRY) values.[19] While it overcomes the limitations of sequential optimization,[14] it also creates a combinatorial explosion of possible designs, making exhaustive screening impractical, even with high-throughput methods.[5] To mitigate this challenge, metabolic engineers often choose to limit the number of pathway elements simultaneously considered[15] and employ the Design−Build−Test−Learn (DBTL) cycle to navigate the vast combinatorial landscape[37] (Figure 1A). In the DBTL cycle, an initial set of designs is introduced into a parent strain (Design phase) and evaluated for TRY performance (Build and Test phases). Insights from these results are used to inform subsequent design iterations (Learn phase), which may use the same parent strain or a higher-performing derivative. Conceptually, this iterative process resembles a Markov decision framework (Figure 1B), where strains represent states, pathway designs are actions, and production outcomes serve as rewards. The overarching goal—

Received: November 18, 2025
Revised: April 7, 2026
Accepted: April 7, 2026
Published: July 22, 2026

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146–3157

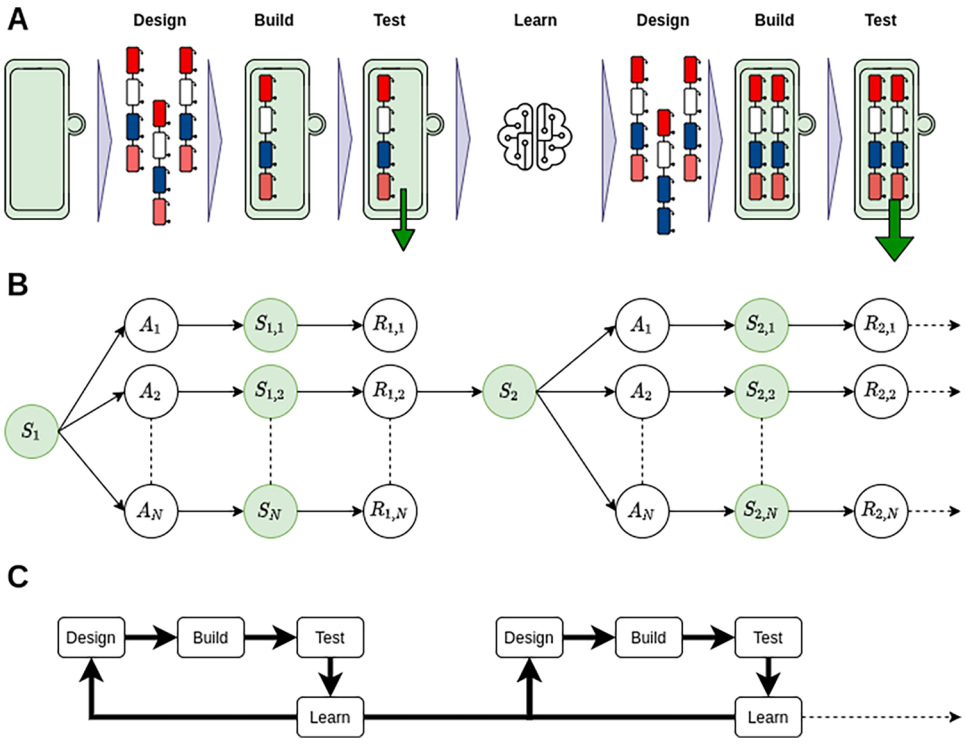


Figure 1. Iterative metabolic engineering from an experimental, computational, and engineering viewpoint. (A) The initial strain undergoes transformation with various pathway designs, integrating these pathways into its genome, resulting in the production of a desired compound (green arrow). By utilizing multiple built strains, a predictive regressor is trained to forecast production outcomes of new designs. This process repeats until strains achieve industrial production levels or reach a theoretical maximum. (B) Iterative metabolic engineering can be framed as a Markov Decision Process (MDP). A selection of actions A (designs) is used to transition to a new state S (new strains), yielding a reward R (production value). Supervised machine learning can be employed to determine the optimal action for a given state S_1, or to proceed further within the MDP model. The role of ML-assisted recommendations is to efficiently navigate this MDP using a model-based policy that chooses which action to take in order to improve the reward. (C) The DBTL cycle in metabolic engineering describes the steps (arrows) involved in advancing through the MDP. In the Learn phase, the experimental results can be used to inform genetic designs based on the current host strain (left arrow from learn to design). Alternatively, new genetic designs can be used on a different host strain that resulted from the previous DBTL cycle (right arrow from learn to design).

maximizing TRY values over successive cycles—parallels reinforcement learning strategies.[44,47]

Machine learning (ML) is increasingly integrated into the DBTL cycle, supporting tasks such as predicting design outcomes,[37] assessing gene importance,[34] and the automated recommendation of new strains.[11,38] Among these, automated design recommendation is particularly promising, as it can close the loop from Learn-to-Design and has demonstrated successful applications.[40,54] However, there remains a limitation in how these machine learning recommendation strategies are applied. First, while advances in genome engineering and high-throughput screening have allowed for targeting many pathway elements simultaneously, the design configuration space often remains limited to a few genes. With configuration spaces on the order of thousands of designs, high-throughput screening and sequencing can achieve relatively dense sampling. This dense sampling of the combinatorial design space may result in increasing production for the recommended strains only marginally, as the strains that were built and used in the training data are already close to optimum of the design space by random chance.[23] Consequently, when the design space is small, the need to balance the exploitation of model predictions with the exploration of new regions — a principle central to Bayesian optimization and reinforcement learning — becomes less important as the design space is already sufficiently explored.

In this study, we examine the impact of substantially expanding the pathway configuration space on design recommendations. Our target is *p*-coumaric acid (pCA), an aromatic amino acid derivative synthesized from phenylalanine and tyrosine, serving as a precursor to numerous bioactive compounds with commercial relevance.[17] We constructed a large library in *Saccharomyces cerevisiae* comprising 18 genes and 20 promoters, yielding approximately 170 million possible configurations. By design, sampling this space is extremely sparse (5×10^{-6}%), even with high-throughput screening, creating conditions where exploration−exploitation trade-offs become critical.[14,53,54] To address this, we implemented a ML-assisted recommendation strategy based on the gradient bandit algorithm,[47] using XGBoost as the predictive model.[3] We experimentally assessed how varying the exploration−exploitation balance influences strain performance and investigated whether the model can predict outcomes beyond its training distribution, a common scenario in iterative optimization when a different parent strain is used. Our findings indicate that an initial large library transformation enables the training of a robust predictive model for subsequent recommendations. Balancing exploration and exploitation improves both gene diversity and pCA production compared to greedy approaches based solely on feature importance.[34] Furthermore, combining a high-performing parent strain with a balanced recommendation strategy that accounts for model uncertainty yields the

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

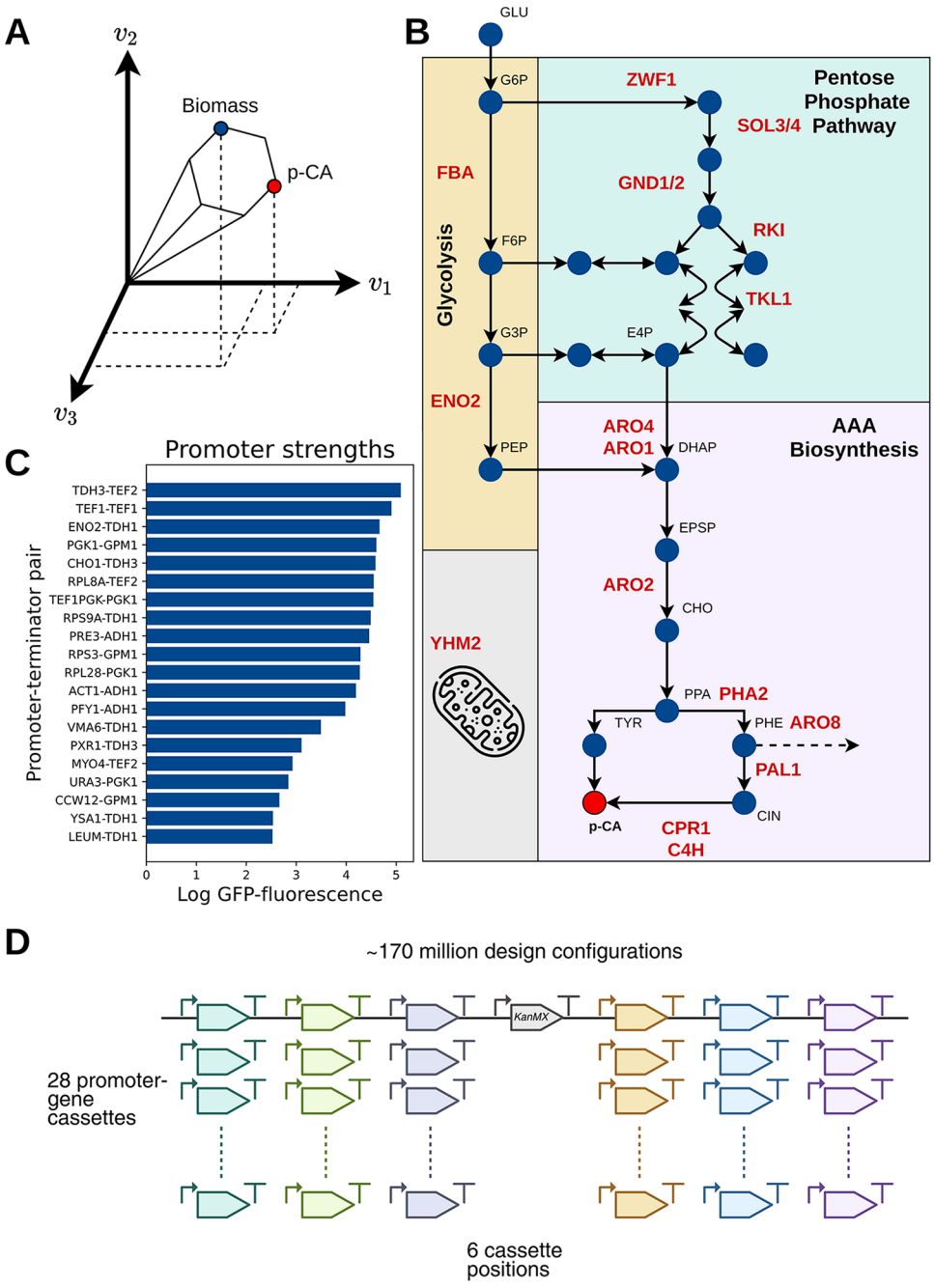


Figure 2. Feature selection using the constraint-based model yeast8. (A) Using the steady-state flux ratio between pCA and biomass as the objective for parsimonious flux balance analysis to identify potential targets ($\frac{v_{\text{Biomass}}}{v_{\text{pCA}}} > 1$) (see Methods). (B) A simplified diagram illustrating the pathway from glucose to pCA, with gene targets identified by pFBA highlighted in red. (C) A collection of already measured promoters for the library.[34] (D) The library contains 28 promoter-gene-terminator cassettes per position and includes 6 engineered positions, with a gray-marked KanMX selection marker. This setup offers a combinatorial action space of roughly 170 million designs, though 18 out of 168 cassette constructions were unsuccessful.

best results. The top strain achieved a 2.37-fold improvement over the previous parent strain,[34] with a titer of 1.23 g/L and a yield of 0.07 g/g on glucose.

RESULTS

Library Design Using the *yeast8* Genome-Scale Model

Supervised machine learning has been used to predict the TRY values of strains,[16,34,40] yet it typically focuses on a limited number of genes and promoters. This restriction weakens the full predictive capabilities that machine learning offers, as strain performance may only slightly surpass the training data. We suggest expanding the combinatorial design space by selecting a wider range of gene targets and promoter values. To select gene targets from the vast array of reactions in yeast metabolism, we utilized parsimonious flux balance analysis (pFBA) on the *yeast8* consensus genome-scale model.[24,29] pFBA seeks to find the minimal total flux through the metabolic networks that maximizes the objective.

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

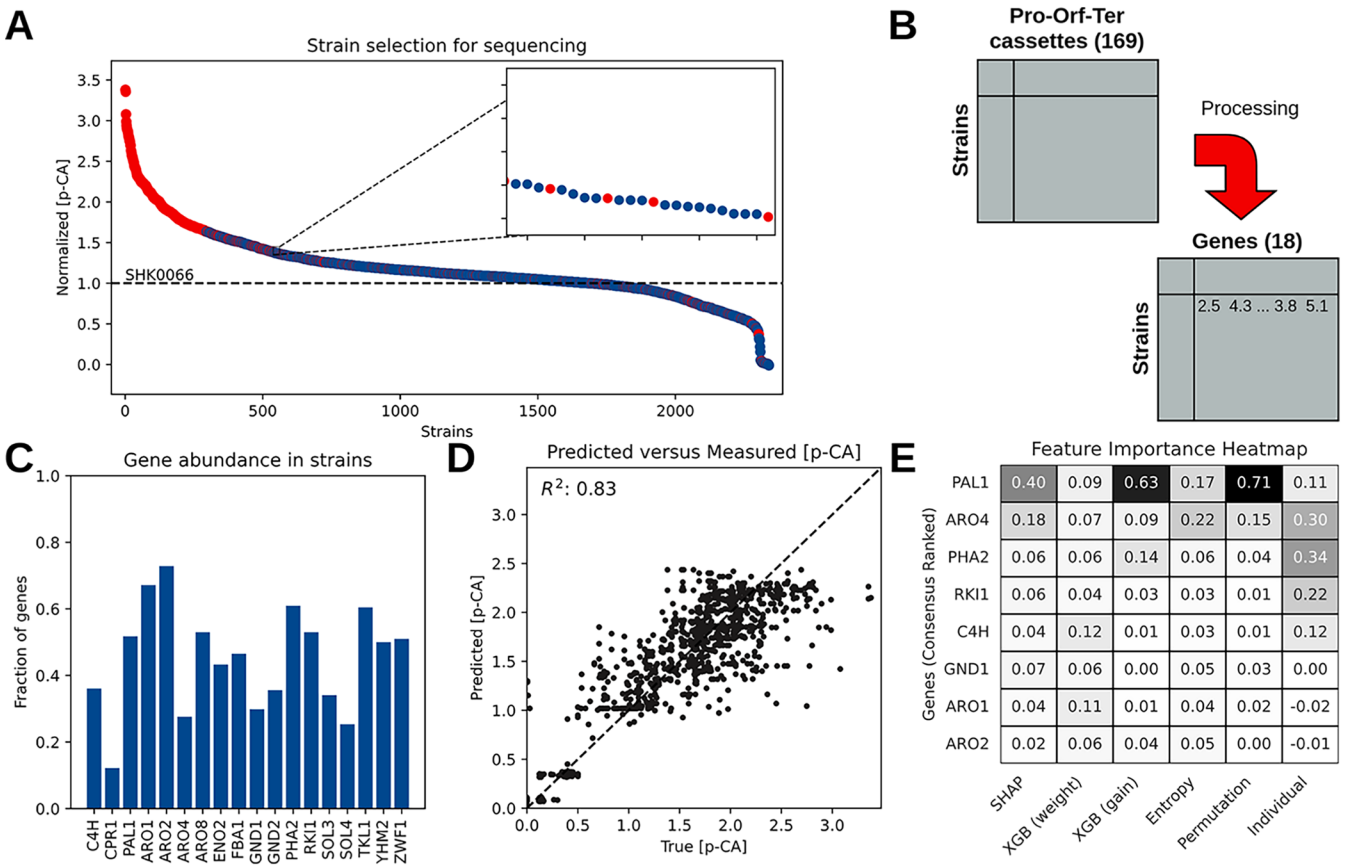


Figure 3. Screening, sequencing, and processing for supervised machine learning. (A) Selection of strains for screening and sequencing. Out of 3000 screened colonies for pCA production, 2350 strains met the quality criteria of the methodological workflow. pCA levels were compared to the parent strain SHK0066, derived from two previous DBTL cycles.[34] From these, six 96-well plates, highlighted in red, were chosen for sequencing, including top producers and a representative sample of lower producers, to minimize machine learning bias using XGBoost.[3] (B) After sequencing and preprocessing (refer to Methods), the cassette count matrix was converted into a gene numeric matrix, quantifying promoter strengths via GFP. (C) Gene abundance in sequenced strains. This fraction indicates gene prevalence across all strains. All library genes were present in the sequenced strains. (D) Predicted vs measured pCA production in test folds during cross-validation. A high Pearson correlation coefficient signifies accurate processing for pCA prediction. (E) Feature importance rankings as determined by six methods (see Methods). The *y*-axis is the consensus gene ranking across these methods.[35] The values in the heatmap are the relative importance per gene.

The selection of genes involved comparing reaction flux ratios between two pFBA objectives, pCA and biomass, identifying reactions that divert flux from biomass to pCA ($\frac{v_{\text{Biomass}}}{v_{\text{pCA}}} > 1$) (Figure 2A). Excluding sink, exchange, and nontargetable reactions, 33 targets were identified. These targets are speculated to shift carbon flux from biomass to the product, suggesting that increasing their protein expression could elevate pCA levels. The targets were validated concerning their impact on pCA, the aromatic amino-acid biosynthesis pathway, and reaction thermodynamics (see Table S1). Consequently, 18 genes were chosen for the DNA library, including two from glycolysis (FBA, ENO2), seven from the pentose phosphate pathway (ZWF1, SOL3/4, GND1/2, RKI, TKL), eight from the aromatic amino-acid biosynthesis pathway (ARO1, ARO4, ARO2, PHA2, ARO8, PAL1, CPR1, C4H), and one involved in oxoglutarate and citrate transfer between the mitochondrion and cytosol (YHM2) (Figure 2B).

Variety in input feature values is crucial for machine learning model predictions. We therefore used 20 promoters, previously quantified by GFP fluorescence, with intensities spanning several orders of magnitude[34] (Figure 2C). This resulted in a library design comprising 28 promoter-gene cassettes, amounting to approximately 170 million design configurations.

Supervised Modeling of the pCA Pathway to Predict Strain Designs

Despite the increasing capabilities of high-throughput screening and sequencing, exhaustively sampling all combinatorial strain designs is both impractical and unnecessary. Effective strain recommendations for subsequent DBTL cycles depend on a reliable supervised model capable of predicting untested designs.[25] To train this model, we transformed a DNA library with a parent strain that resulted from a previous pCA optimization study,[34] randomly picked 3000 colonies, and screened their pCA production using Nuclear Magnetic Resonance (NMR). The pCA production was adjusted and normalized against the parent strain (SHK0066).[34] There was a wide variance in pCA production among the strains, with 5% producing over double the amount of SHK0066 and 27% producing less (Figure 3A). To verify the performance of the highest producers, we rescreened the top 86 strains with four biological replicates, resulting in more conservative estimates; though they remained high producers (see Figure S3A). These remeasured top producers and data from a previous DBTL cycle were used in the training data.[34]

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

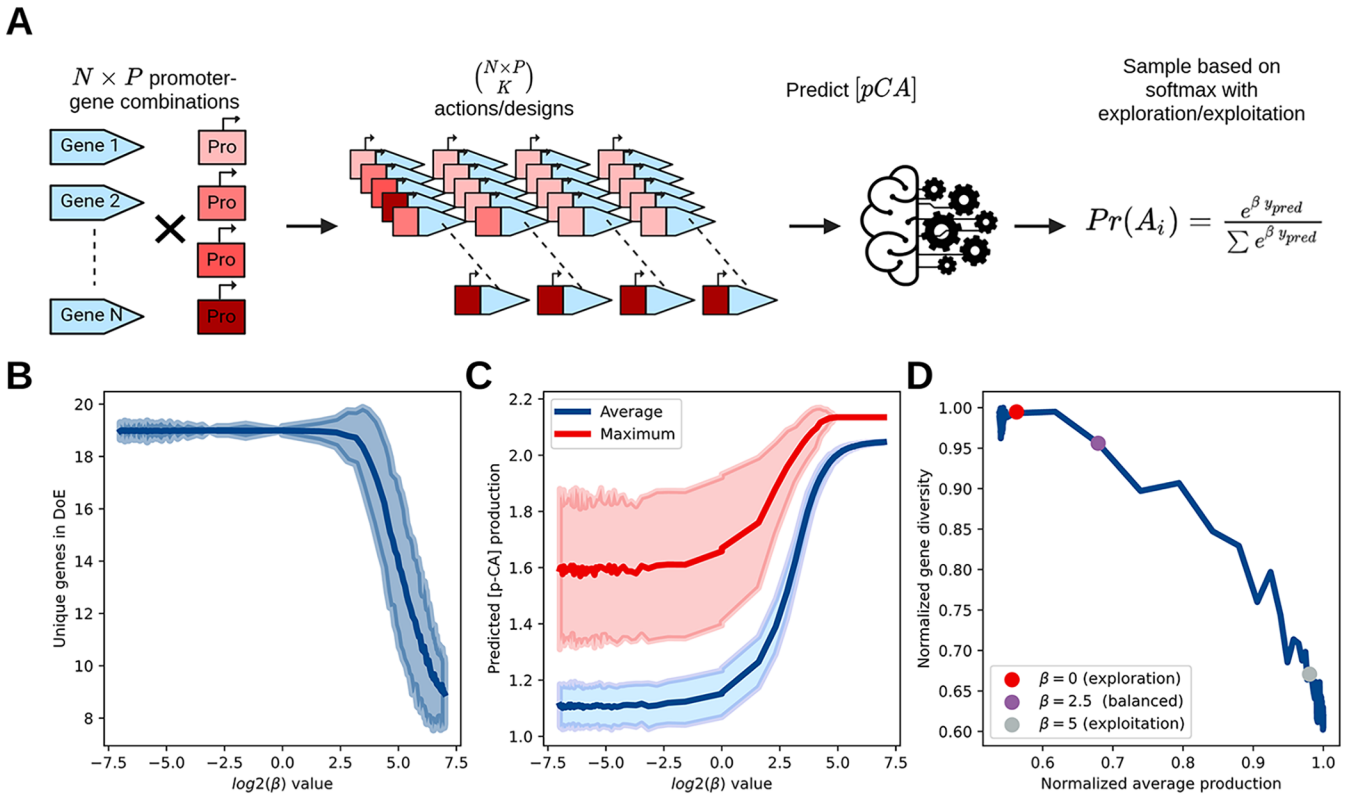


Figure 4. Exploring machine learning-enhanced automated recommendation with a balance between exploration and exploitation. (A) Overview of the strategy: four promoters (P) and 18 genes (N) form a set based on pathway positions (K) to create an action space. This model forecasts pCA outcomes, with designs sampled from a softmax distribution featuring an exploration-exploitation parameter (β). (B) Impact of exploration/exploitation on design diversity. For each β value, 25 samples were drawn, repeated 100 times (shaded region). Increasing β reduces gene set diversity. (C) Average and maximum predicted production relative to β. More exploitation leads to higher average predicted production. (D) Balance between gene diversity and average production. Three β settings were selected for experimental validation of the recommendation algorithm and to study the effect of the exploration-exploitation trade-off in practice.

After a library transformation, the precise genetic makeup of the strain remains unknown after NMR screening. To address this, we selected a representative sample across the production spectrum and conducted whole genome sequencing to identify integrated library cassettes (see Methods) (Figure 3A). We further processed the copy count matrix by aggregating cassettes sharing identical genes into numerical values based on promoter strengths, effectively reducing the data set's dimensionality from 169 library cassettes to 18 genes (Figure 3B). Our findings revealed that, aside from CPR1, all genes were prevalent in the sequenced strains, exceeding 10% abundance (Figure 3C). This observation aligns with CPR1's limited inclusion in the library design, where it appeared in only five out of 150 successful cassettes. Despite the library's intention to introduce six genes per strain, unexpected homologous recombination led to strains with over 20 introduced genes (see Discussion). Although these strains are likely genetically unstable, they could still provide valuable information for training the machine learning model as they result in strains that were not purposefully designed. For example, a strain with many integrated genes of the same copy could have higher expression levels than could be purposefully be designed. We therefore included these strains in the data set.

For modeling, we employed the ensemble method XGBoost with a density-based weighted loss function, emphasizing underrepresented data points via kernel density estimation[46] (see Methods). Model predictions were validated through cross-validation, showing a high Pearson correlation (PCC = 0.83) between predicted and actual pCA production on the validation sets, indicating robust predictive performance. Despite this, there was some under-prediction of high-yield strains and overprediction of low-yield strains (Figure 3D). To further verify the model, we identified genes that contribute to its performance and enhance pCA production. Six feature importance methods were used to rank genes (see Methods) for the XGBoost model. A Borda consensus ranking of the first 8 genes is reported in Figure 3E.[35] The genes PAL1, ARO4, and PHA2 were highlighted as the top genes that have a strong influence on pCA, despite observed variability across feature importance rankings (see Figure S4). These genes were previously found to boost pCA production in *S. cerevisiae*,[13,32,34] supporting the model's reliability for further application in recommending strains.

A Model-based Exploration-Exploitation Strategy for Strain Recommendation

To improve pCA production using the XGBoost model, it is important to implement a strain recommendation strategy suitable for the complex combinatorial design landscape. This strategy should create a practical action space anticipated by the model from a metabolic engineering viewpoint (Figure 1B) and address the uncertainty of model predictions. Although the model performs well in cross-validation (Figure 3D), its accuracy often declines with new strain designs due to biological and technical variability in pCA concentration,

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

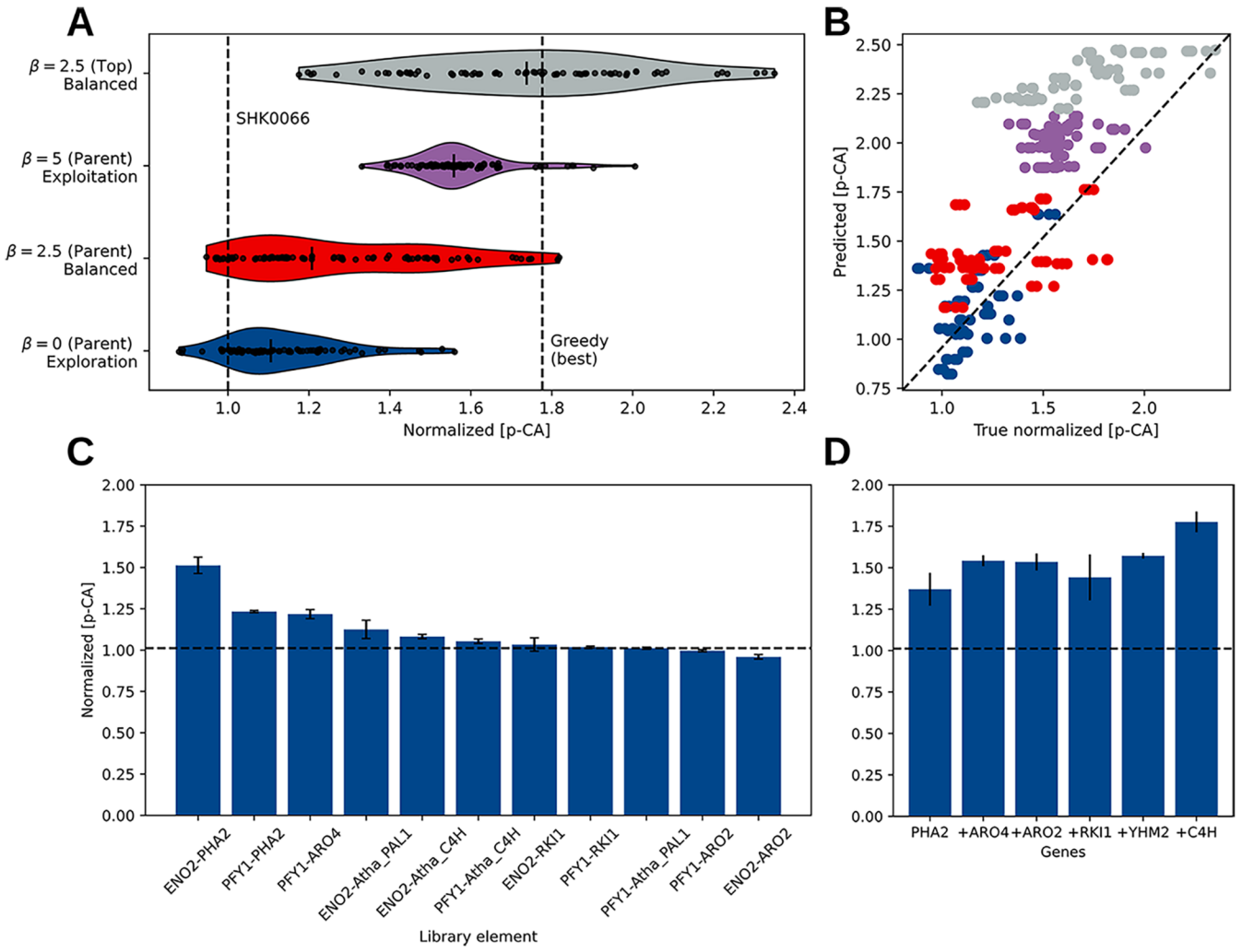


Figure 5. Experimental validation of the impact of balancing exploration and exploitation in extensive combinatorial spaces. (A) Scenarios characterized by exploration, balance, and exploitation (β = [0, 2.5, 5]) were tested using parent strain SHK0066. The values for pCA production are presented relative to the production by SHK0066.[34] A balanced scenario was evaluated for the top predicted strain SHK0073. (B) A comparison of predicted versus actual [pCA] production across the four scenarios. The XGboost algorithm effectively predicted the general trend, although it underestimated the variability in actual [pCA] production. (C) The experimental impact of individual genes included in the library design was assessed. Promoters of varying strength, PFY1 (weak) and ENO2 (strong), were utilized. The genes PHA2, ARO4, C4H, and PAL1 significantly improved pCA production (Wilcoxon rank sum test, $p < 0.05$). (D) Construction of strains using a greedy approach based on the model's best-predicted strain. Six strains were developed with an increasing number of cassettes predicted to maximize production.

particularly in high-performing strains (see Figure S3A), and possible overfitting. More fundamentally, this decrease is due to the fact that new parent strains, which are outside the model's training data, are often used in later DBTL cycles. Thus, balancing model prediction exploitation with design space exploration for better returns is key.[47] We propose a ML-assisted recommendation strategy that effectively balances exploration and exploitation to successfully optimize pathways.

We create a design space by combining four promoters with 18 genes, resulting in 72 different cassettes (Figure 4A). For each pathway design, four gene cassettes are selected (see Methods). The predicted pCA production for these *in silico* strain designs is determined using the aforementioned model. A softmax distribution is applied to sample actions, giving each action a sampling probability based on the model's predicted pCA production. The exploration-exploitation parameter β in the formula of Figure 4A influences the sampling bias: a higher β increases the bias toward model-driven recommendations for high-yield strains. To illustrate the impact of increasing β, 25 strains were sampled from this softmax distribution with varying β values. As the exploitation increases, the number of unique genes decreases (Figure 4B), while both average and maximum strain performance improve (Figure 4C). This indicates that strains sampled in highly exploitative contexts become more similar, which can be problematic if the supervised model has biases, potentially missing more optimal producing strains. There is a clear trade-off between the genetic diversity of sampled strains and their average predicted production (Figure 4D).

Experimental Validation Reveals a Successful Automated Recommendation Strategy and Importance of Balancing Exploration and Exploitation

The exploration-exploitation trade-off is a key concept in Bayesian optimization and reinforcement learning.[8,47] This trade-off holds particular importance in the recommendation strategy previously discussed and is also relevant in metabolic engineering;[11,40,44] yet, experimental validation on a consistent

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

optimization target remains sparse. To evaluate its role in effective recommendation strategies, we examined four scenarios with varying degrees of exploitiveness. For SHK0066, this consists of an explorative, balanced, and exploitative scenario, with β values set at 0, 2.5, and 5, respectively (Figure 4D). A fourth scenario involved choosing the best predicted strain, SHK0073, with four integrated genes, as the parent strain using a balanced approach.

In each scenario, 25 strains were constructed with four biological replicates (Figure 5A). The dashed lines indicate the parent SHK0066 and the top-performing strain selected via a greedy recommendation method for the six genes included (see Methods). The average production of strains using parent SHK0066 rises with increased exploitative approaches. Several strains in both the balanced and exploitative scenarios outperform the greedy recommendation strategy, highlighting the need for caution when relying solely on predictive models greedily. The balanced scenario exhibits the greatest variance, while the exploitation scenario shows the least variance in pCA production. This variance in the balanced scenario is anticipated: strains with greater genetic diversity display a wider range of pCA production compared to strains with similar genetics. These findings confirm the trade-off inherent in this recommendation strategy.

In the final scenario, using the top producer SHK0073 with four genes included, the average pCA production reached even higher levels, with several strains significantly outperforming the greedy recommendation strategy. The best strain achieved production levels 2.37 times greater than SHK0066, 1.33 times more than the greedy strategy (which included six genes), and 1.18 times more than the best strain from the initial library transformation round (which included ten genes via homologous recombination) (Figure S3B). Note that there is an even higher variance in production than in the balanced scenario of SHK0066.

The higher variation in production performance for this scenario could be explained in two ways: (1) high variation is generally expected in this balanced scenario, and (2) the model was not specifically trained on this host strain. Notably, production variance was higher than that in the SHK0066 balanced scenario. This latter point highlights the importance of incorporating exploration into the approach: models trained on previous strains may not perform well out-of-distribution, a common limitation in machine learning. Therefore, addressing this uncertainty is crucial when using automated recommendation strategies.

Deeper Validation of Predictive-ness and Feature Importance

While the outlined recommendation strategy seems effective, questions persist about the model's predictive performance and explainability. High predictive performance is crucial because it enables a more exploitative recommendation approach. Explainability matters for understanding biological processes and ensuring the model's learning is biologically grounded.[26]

When validating the model's predictive performance for pCA production, we find that the overall trend in the scenarios is accurately predicted (Figure 5B). This is reflected in a strong correlation between predicted and measured pCA concentration across all scenarios (Table 1). However, large differences in correlation are observed for the individual scenarios. For the balanced exploration-exploitation scenario with parent strain SHK0073, a high correlation is observed.

Table 1. Correlations of the Predicted versus true pCA Production of Figure 4B

scenario	Pearson correlation	Spearman rank correlation
all	0.63	0.78
β = 0 (SHK0066)	0.28	0.43
β = 2.5 (SHK0066)	0.15	0.32
β = 5 (SHK0066)	0.02	0.16
β = 2.5 (SHK0073)	0.55	0.76

For the scenarios with parent strain SHK0066, the correlation diminishes with increasing exploitativeness. This finding can likely be explained by two factors. First, while the trained XGBoost model may capture the overall distribution well, it might fail to predict more fine-grained differences between strains accurately. Second, measurement noise has a more dominant impact on the correlation coefficients for scenarios where the pCA variance is low. These results suggest that the model generally captures behavior well, but higher variation from exploratory designs is crucial to avoid overfitting or overestimation.

We conducted two experiments to determine whether the feature importance results from the model could be experimentally validated. In the first experiment, individual promoter-gene cassettes were integrated into SHK0066 to verify the feature importance rankings shown in Figure 3E. Genes such as PHA2, ARO4, PAL1, and C4H positively influenced pCA production (t test, p-value 0.05) (Figure 5C) and were identified by at least two feature importance methods. Although the rankings differed by method, this suggests that the model's learned representation aligns with biological reality. The second experiment involved a greedy recommendation approach, building the best predicted strain by incrementally adding genes (Figure 5D). Despite observing increased production with more genes, the model overestimated this improvement. The top strain from this method produced less pCA than several strains suggested by our recommendation algorithm, indicating that caution is needed with purely greedy strain construction.

DISCUSSION

This study examines the potential of ML-assisted recommendation tools in metabolic engineering to navigate the vast combinatorial space that cannot be explored experimentally. As the combinatorial design space expands, the focus shifts from pure exploitation of model predictions toward a more deliberate balance between exploration and exploitation. Our findings suggest that recommendation strategies incorporating both exploration and exploitation outperform purely greedy approaches that rely exclusively on model prediction (Figure 5A, dashed line). Importantly, the optimal degree of exploitation appears to depend strongly on the predictive accuracy and certainty of the ML model. In the case of the parent strain SHK0066, where predictive performance was relatively strong, a more exploitative strategy identified the highest pCA-producing strains. However, when model predictions are uncertain or less reliable, heavily exploitative strategies risk converging on local optima, limiting the discovery of superior strain designs elsewhere in the design space. In such scenarios, a balanced exploration−exploitation strategy could become advantageous, as it increases the probability of identifying globally optimal solutions. This is particularly relevant when transitioning to new parent strains,

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

where model uncertainty may be greater. Under these conditions, a balanced approach represents a more robust and risk-mitigating strategy for strain optimization.

The combinatorial design space in this study was constructed from 18 genes and 20 promoters, arranged across six library positions with twenty-eight cassettes per position. The targeted genes were identified through a combination of pFBA-based feature selection and additional filtering using using database resources and existing literature. This led to the inclusion of several previously characterized genes.[27] While this approach ensured biological relevance, it may bias target selection toward well-studied genes. Future studies could benefit from more data-driven gene selection strategies that are less dependent on prior knowledge, potentially enabling the discovery of novel targets and further improving metabolic pathway optimization.[1] For the promoters, we selected 20 unique promoters that were quantified by GFP fluorescence in previous work.[34] The large set of promoters was chosen for two reasons: (i) to reduce the likelihood of homologous recombination by assigning distinct promoters to each position, and (ii) to introduce varying expression levels. This increases the informational content of the 18 genes for modeling, despite further inflating the combinatorial space. Although these measures aimed to minimize recombination, such events were still anticipated due to the repetitive nature of the library design. Indeed, some strains unexpectedly contained more than ten genes instead of the intended six, resulting in genetic instability. For example, the two top-performing strains from the first DBTL cycle integrated 26 and 23 genes (including KanMX markers), respectively (Figure S3B). Despite their instability, these strains provided valuable insights and were included in XGBoost model training.

It is important to note that while GFP fluorescence serves as a convenient proxy for expression level, the magnitude of a promoter's effect can differ between targeted enzymes. Similarly, the expression strength may be influenced by the DNA library position of the integrated promoter. In our modeling framework, we assumed that for a given gene, promoter strength and enzyme expression are related by a monotonically increasing function. This assumption proved sufficient to achieve solid predictive performance, likely due to the inherent flexibility of the supervised machine learning model. Nevertheless, future work could benefit from directly linking promoter strengths to proteomic measurements, which may better capture enzyme-specific expression dynamics and further improve model accuracy.[40]

A key aspect of this work is the balance between exploration and exploitation within the combinatorial design space. Although the importance of this effect has been noted in prior work,[2,38,40] consistent experimental validation of the exploration-exploitation trade-off has been only limitedly explored in the metabolic engineering literature.[54] While the exploitative scenario ($\beta = 5$) produced high-performing strains, it did so at the cost of genetic diversity among the proposed designs. This reduction in diversity may lead to suboptimal decisions when model performance is lower than that observed for the pCA model used here. Consequently, accounting for model uncertainty is essential when determining the appropriate balance between exploration and exploitation. Although we demonstrate the significance of this balance, we do not prescribe a specific method for achieving it (see Figure S5). One potential approach could involve maximizing the product of genetic diversity and average production, as illustrated in Figure 4D, similar to the strategy employed by the MODIFY algorithm for directed evolution.[6]

For the second round, we selected specific strains recommended by the algorithm. This resulted in strains with a pCA titer of 1.23 g/L and a yield of 0.07g/g glucose, which was a 2.37 times improvement over the original parent strain SHK0066.[34] While higher pCA production has been reported,[28] the strain designs reported in this study may serve as a basis for further improvement to enable industrial use.[33] This would include the need for incorporating bioprocess conditions in the model formulation. We expect that culture and process conditions can be incorporated using mechanistic modeling approaches, such that they generalize better during bioprocess scale-up than pure machine learning approaches.[21]

As the methodology for constructing the strains in the second round differs from library transformation, the number of strains that can be built is substantially smaller than what is achievable through a library-based approach using comparable experimental resources. However, this recommendation strategy could also be adapted for library design using the approach illustrated in Figure 4A. An additional step would involve generating a probability distribution for each library element at each position. Experimentally, this can be implemented by adjusting the concentration of library cassettes, thereby modulating the probability of integration during library transformation. We experimentally verified that cassette concentration can influence integration likelihood (see Figure S2). This approach would enable larger-scale follow-up rounds.

To evaluate feature importance methods, we experimentally determined the individual contribution of specific genes to p-CA production (Figure 5C). Three genes—PHA2, ARO4, and PAL1—emerged as key contributors from the consensus of all feature importance approaches.[35] PHA2 catalyzes the conversion of prephenate to phenylpyruvate and has previously been reported to significantly influence pCA production.[28,31] AR04, which catalyzes the first step in the shikimate pathway, is known to be critical for the parent strain SHK0066 used in this study and in other pCA yeast cell factories.[34,42] PAL1, originating from *Arabidopsis thaliana*, offers an alternative to the TAL-route[28,42] and was also included in the parent strain.[34] For the other considered genes, more variability between feature importance methods for gene importance rankings was observed (Figure S4). For these additional candidates, some of which were experimentally verified, their effects on pCA production were also shown to be less consistent (Figure 3E). Some genes showed a modest positive effect (e.g., YHM2), whereas others exhibited no measurable improvement or even a slight negative impact on production (e.g., ARO2 and RKI1). These results indicate that feature importance rankings may not always correspond to improvements in production.

METHODS

Organisms and Media

S. cerevisiae strains originate from CEN.PK113-7D. *p*-Coumaric (pCA) production strain BMP+PAL-C4H-CPR (SHK0066) from study[34] was used as starter strain. Strains were grown at 30 °C in Yeast Extract Phytone Dextrose media (YEPhD, 2% Difco phytone peptone (Becton-Dickinson (BD), Franklin Lakes, NJ, USA), 1% Bacto Yeast extract (BD) and 2% D-glucose (Sigma-Aldrich, St Louis, MO, USA)). When required, antibiotics were added to the media at appropriate concentrations: 200 μg/mL nourseothricin (Jena

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

Bioscience, Germany), 200 μg/mL Geneticin (G418, Sigma-Aldrich). *E. coli* DH10B (New England BioLabs, Ipswich, MA, USA) was used as cloning strain and grown at 37 °C in 2*Peptone Yeast Extract media (2*PY, 1.6% tryptone peptone (BD), 1% Bacto yeast extract (BD) and 0.5% NaCl (Sigma-Aldrich)). When required, antibiotics were added to the media at appropriate concentrations: 100 μg/mL ampicillin (Sigma-Aldrich), 50 μg/mL neomycin (Sigma-Aldrich). Solid medium was prepared by addition of Difco granulated agar (BD) to the medium to a final concentration of 1.5% (w/v).

Cassette Construction

Gene Targets. To determine which gene targets should be included in the library, parsimonious flux balance analysis (pFBA) is performed[24] on the genome-scale model Yeast8.[29] We use the pFBA solution for a pCA objective and the biomass objective to calculate a flux ratio. This flux ratio indicates which fluxes are rerouted when diverting carbon flux from biomass toward pCA and was used to select genes for further validation ($\frac{v_{Biomass}}{v_{pCA}} > 1$). After excluding sink, exchange, and nontargetable reactions, 33 potentially interesting gene targets remain.

We further aimed to substantiate gene targets by performing literature enrichment analysis using the STRING database,[48] thermodynamic information from the equilibrator database,[36] and inhibitor information from the BRENDA database.[45] Finally, we assess the importance of each gene through a literature review (see Table S1). This resulted in a set of 19 validated genes that are used in the library. Open reading frames (ORF) were codon pair optimized for *S. Cerevisiae*.[43] Cassettes were designed as described in rer 50 and constructed as described in ref 34.[34]

Promoters. A set of 20 previously established promoters, with varying strengths quantified through GFP fluorescence, was used in the library design.[34] The expression strength of these promoters spans several orders of magnitude and can be used to regulate gene expression. We assume that the expression pattern might differ between genes with the same promoter, but that the relative expression per gene is similar.

Library Design. The number of positions in the library was limited to six factors to ensure sufficient transformation efficiency. To avoid excluding certain gene combinations during transformation, we decided that for each position in the construct, all genes should be present with two distinct promoters.[34] For genes that were previously considered in the reference strain (C4H, CPR1, PAL1, ARO4), we chose one promoter per library position. This library consists of 168 cassettes, of which 150 were successfully built. While incorporating all genes ensures that we do not exclude certain combinations, it might lead to increased homologous recombination events during transformation. The library design can be found in the Supporting Information.

S. cerevisiae Strain Construction

Transformation. Strains were constructed as described in ref 34. In short, host strain SHK0066, pre-expressing Cas9 was used for CRISPR mediated integration of the cassette library into the target locus INT28, see supplementary info for gRNA design and genome locus position. Cassette libraries were prepared with equimolar amounts of cassettes (100 ng/kb). Per cassette library previously considered genes (C4H, CPR1, PAL1, ARO4) were added with higher dosage as follows; library 1:300 ng/kb, library 2:200 ng/kb, library 3:100 ng/kb, library 4:50 ng/kb, the library design can be found in the Supporting Information. Linear gRNA (210 ng/kb) and linear backbone with homologous sites to the gRNA (35 ng/kb) were transformed together with the cassette libraries into the host strain following LiAC/ssDNA/PEG method.[9]

For design-based transformation, the cassettes were prepared per pathway design in equimolar amounts and transformed with the linear gRNA and linear backbone following the same LiAC/ssDNA/PEG method. The cassette design with on five prime and three prime side 50 base- pair connector sequences homologous to INT28 facilitate in vivo recombination of cassettes into the genome.

The transformed cells were plated on a Qtray (Nunc) containing yephD agar and appropriate antibiotics. 48-dividers (Genetix) were used in the Qtray when a pathway design was done. Qtrays were incubated for 3 days at 30 °C and picked with Qpix 450 (Molecular Devices) into 96 MTP (Nunc) containing YephD agar with appropriate antibiotics.

pCA Screening in *S. cerevisiae*

Strain Fermentation. pCA production experiments were conducted as follows; colonies were grown in a microtiter plate (MTP) containing YephD agar medium and incubated for 48 h at 30 °C, 750 rpm, and 80% humidity. Cultures were inoculated in a half deep well plate (HDWP, Thermo Fisher Scientific, AB-1127) containing 350 μL YephD and appropriate antibiotics as a preculture. Strains were reinoculated in main culture into HDWP containing 350ul Verduyn Luttik with 2% glucose[4] and incubated for 48 h at 30 °C, 750 rpm, 80% humidity. Every plate contains a medium control and 16 times the parent strain SHK0066 for later normalization.

pCA Measurement with NMR. For Flow-NMR measurements, 250 μL of the end of the fermentation broth was sampled into a 96-deepwell plate and mixed with 500 μL of acetonitrile using the EPmotion (Eppendorf) to initiate cell lysis. The samples were centrifuged at 4000 rpm for 10 min, and 500 μL of supernatant was sampled into a new deep well plate. 70% of the liquid was evaporated, and 100 μL of 1 g/L internal standard 1,1-difluoro-1-trimethylsilanyl methylphosphoric acid (FSP, Bridge Organics) was added. The samples were lyophilized to remove nondeuterated solvents. To the lyophilized samples, 600 μL of D2O (Cambridge Isotope Laboratories (DLM-4)) was added and homogenized. The samples were analyzed on a CTC PAL3 Dual-Head Robot RTC/RSI 160 cm robotic autosampler (CTC Analytics AG, Zwingen, Switzerland) fluidically coupled to a Bruker spectrometer Avance III HD 500 MHz UltraShield [26]. 1H spectra were recorded with standard pulse program (zgcppr) with following parameters: 16 scans, 2 dummy scans, 33k data points, 16.4 ppm spectral width, 1.2 s relaxation delay (d1), 8 μs 90° pulse, 2s acquisition time, 15 Hz water suppression, and fixed receiver gain (rg) of 64 Spectra were processed and analyzed using Topspin 4.1.4 (Bruker). Spectral phasing was applied and spectra were aligned to 3-(trimethylsilyl)-1-propanesulfonic acid-d6 sodium salt (DSS-d6, Sigma-Aldrich) at 0 ppm. Auto baseline correction was applied on the full spectrum width. Additional third-order polynomial baseline correction for selected regions was applied if needed. The amount of pCA (doublet, 6.38 ppm, n = 2H) was calculated relative to the signal of FSP.[34]

Strain Quality Control Using Whole Genome Sequencing

Choice of Screened Strains for Sequencing. Based on the screening of 3000 strains, we chose a representative set of strains for sequencing. We chose three 96-wells plates for the top producers, two 96-wells plates for the strains that showed improved pCA production but that are not in the top, and one 96-wells plate for strains that performed worse than the parent strain.

Whole Genome Sequencing. *S. cerevisiae* cells (OD600 of 5-10) were pelleted and lysed as follows; 140 mg Zirconia beads (0.5 mm diameter, Biospec products) and 400 μL ZymoBIOMICS Lysis buffer (Zymo Research) were added to the cell pellet. Cells were lysed for 5 min at 1500 rpm with the Geno/Grinder 2010 (Cole-Parmer). gDNA was further isolated following ZymoBIOMICS 96 DNA kit (Zymo research) according to manufacturers protocol. The isolated genomic DNA was quantified using the Tapestation 4200 (Agilent) with a genomic DNA screentape (Agilent) and Nanodrop (Thermofisher Scientific). Genomic DNA was sequenced per strain or per library pool of transformants using the native barcoding kit (SQK-NBD114.96) from Oxford Nanopore Technologies on a Promethion device (FLO-PRO114M) according to manufacturer's protocol. For pool sequencing superaccuracy basecalling was used and for strain sequencing high accuracy basecalling.

Data Processing of Individual Clones. DNA sequencing reads obtained from picked individual colonies after transformation, were cleaned using Porechop.[51] Cleaned reads were then assembled into genomes using Canu,[20] with the genome size parameter set to 12

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

Mbp, and were followed by one round of polishing by Racon[49] and Medaka.[52] To identify where the inserted BioBricks were located, annotations of the cassette sequences (such as promoter, open-reading-frame, terminator annotations) were transferred to the genome assemblies by blast-based sequence alignment, leading to genome assembly annotations in GFF format. Cassette annotations were only transferred to the genome assemblies if at least 95% of the cassette annotation was included in an alignment hit to the genome, with at least 95% identity.

Machine Learning-Assisted Recommendation Strategy

Ensemble learning tree methods have been shown to perform well on combinatorial pathway optimization data.[23,38] We therefore trained an XGBoost,[3] a regularized gradient boosting algorithm, with a modified loss function that is more specific to our purpose.

Density-Based Weighting of Loss Function. Machine learning methods typically assume that data is balanced; that is, the target variable is uniformly distributed. However, in production screening data, this is often not the case. High performing strains are typically underrepresented, while strains with minor increases in performance are overrepresented. To mitigate the effect of this imbalanced distribution on the mean squared error (MSE),[41] we used DenseWeight.[46] DenseWeight weighs the influence of data instances on the loss function according to the target variable frequencies estimated through kernel density estimation. DenseWeight requires setting two hyperparameters that determine how much the density estimation should influence the weighting, which were set to $\alpha = 0.5$ and $\epsilon = 0.01$.

XGboost requires the calculation of the gradient and Hessian of the loss function for efficient training, and due to the modified loss function, it needs to be adjusted. The modified loss function (L_{DensLoss}) is the mean squared error weighted by a function ($f_w(y_i)$) that is dependent on the target variable density and its hyperparameters. The loss function, Jacobian, and Hessian with respect to a data instance y_i are then given by (eqs 1−3).

$$L_{\text{DenseLoss}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} f_w(y_i)\cdot \text{MSE}(y_{\text{pred}}, y_i) \tag{1}$$

$$\frac{\partial L_{\text{DenseLoss}}}{\partial y_i} = 2(y_{\text{pred}} - y_i)\cdot f_w(y_i) \tag{2}$$

$$\frac{\partial^2 L_{\text{DenseLoss}}}{\partial y_i^2} = 2\cdot f_w(y_i) \tag{3}$$

XGboost Training. Bayesian hyperparameter optimization was performed on the training data using scikit-optimize.[12] This resulted in the following hyperparameters (*reg_lambda = 1, max_depth = 2, tree_method = "auto"*). The customized loss function was used as described above. Two hundred rounds of boosting were performed with an early stopping criterion of 40. This means that if the validation mean squared error does not improve after 40 rounds, training is stopped. Model validation was performed on a ten percent random split using sklearn and was evaluated using the Pearson correlation coefficient between the true and predicted values. We further validated the model using an external validation set of designs (see below).

Recommending New Strains Using Gradient Bandit Principles. To recommend new strains for the second round, we start by defining the action space (Figure 1B). In the case of $N = 20$ genes with $P = 4$ cassette positions, we choose $X = 4$ different levels of promoter strength, which result in $\binom{NX}{P} = 1028790$ actions. Actions that have multiple integrations of the same gene are filtered out. Then, XGBoost is used to predict the potential outcome of the set of actions given the reference strain. We chose two reference strains: the parent of the first library transformation round (SHK0066) and the best predicted action strain for four pathway positions.

As a recommendation strategy based on predicted outcomes, we used a gradient bandit approach[47] using the softmax distribution with an exploration/exploitation (β) parameter (eq 4):

$$Pr(A_i) = \frac{e^{\beta y_i}}{\sum e^{\beta y_i}} \tag{4}$$

The exploration/exploitation parameter β is an important parameter for automated recommendation and represents a trade-off between the diversity of the set of chosen designs in terms of gene content and the expected performance of the strains in terms of production.[6] Here, we have tested three exploration/exploitation values ($\beta = 0$, $\beta = 2.5$, $\beta = 5$) to explore this trade-off in more detail.

Feature Importance Methods. Six feature importance methods were used to retrieve gene importance rankings from the trained model. Two methods are intrinsic to the XGBoost model: the *weight* method ranks features based on the number of times a feature is used to split data across the trees, and the *gain* method bases this on the average information gain across all splits in which the feature is used. The three other methods are model independent. One popular method for explaining machine learning models is the SHAP values.[30] We also tested feature importance ranking based on the area under the curve of gene-promoter threshold plots, as used in ref 23. Furthermore, a permutation importance method from sklearn was used.[39] Finally, the individual contribution of each gene was calculated by predicting the outcome of a single promoter-gene perturbation on the trained model.

■ ASSOCIATED CONTENT

Data Availability Statement

All generated data is accessible for future research. This comprises DNA sequences of the strains from the initial DBTL cycle, NMR screening results of the 3000 strains, and processed data sets for machine learning. These resources are available in the 4tu repository (DOI: 10.4121/fa0782ab-760d-4fa7-babf-09bdaab0f509). The related code is accessible on GitHub https://github.com/AbeelLab/ml-assisted-p-coumaric-acid-optimization.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.5c00864.

DNA sequencing files can be found in the 4tu-repository. Code and processed files required for reproducing this work can be found in the GitHub repository (see Code and data availability) (XLSX)

List of promoters, ORFs, and terminator sequences used in the study: integration sites used for strain construction; the list of strains used in this study and their genotypes; the list of plasmids used in this study; the list of primers used in this study; and sgRNA design (XLSX)

Library design files for DBTL round 1 and the follow-up machine-learning-assisted recommendations (XLSX)

Additional information on parent strain lineage, literature validation of chosen gene targets, bar plots of feature importance rankings, whole genome sequencing library pool analysis, and rescreening of the top producers from the first DBTL cycle, as well as a rebuild of the top 2 producers (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Thomas Abeel − *Intelligent Systems, Delft University of Technology, 2628 AT Delft, The Netherlands; Infectious*

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157

Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge 02142, United States; orcid.org/0000-0002-7205-7431; Email: t.abeel@tudelft.nl

Authors

Paul van Lent − *Intelligent Systems, Delft University of Technology, 2628 AT Delft, The Netherlands;* orcid.org/0009-0001-2887-0193

Rianne van der Hoek − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Sara Moreno Paz − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Irsan Kooi − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Moniek Jonkers − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Priscilla Zwartjens − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Joep Schmitz − *Department of Science and Research, dsm-firmenich, 2613 AX Delft, The Netherlands*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.5c00864

Notes

The authors declare the following competing financial interest(s): J.S., R.v.d.H., S.M.P., I.K., M.J., and P.Z. were employed by dsm-firmenich at the time of the study.

■ ACKNOWLEDGMENTS

The authors thank the AI4b.io consortium and the Delft Bioinformatics Lab for their relevant discussions during progress meetings. This work is supported by the AI4b.io program, a collaboration between TU Delft and DSM-Firmenich, and is fully funded by DSM-Firmenich and the RVO (Rijksdienst voor Ondernemend Nederland).

■ REFERENCES

(1) Alonso-Gutierrez, J.; Kim, E. M.; Batth, T. S.; et al. Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng.* **2015**, *28*, 123−133.

(2) Borkowski, O.; Koch, M.; Zettor, A.; et al. Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat. Commun.* **2020**, *11*, No. 1872.

(3) Chen, T.; Guestrin, C.et al.Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining* 2016; pp 785−794.

(4) Dekker, W. J. C.; Wiersma, S. J.; Bouwknegt, J.; et al. Anaerobic growth of *Saccharomyces cerevisiae* CEN. PK113−7D does not depend on synthesis or supplementation of unsaturated fatty acids. *FEMS Yeast Res.* **2019**, *19*, No. foz060.

(5) Dietrich, J. A.; McKee, A. E.; Keasling, J. D. High-throughput metabolic engineering: advances in small-molecule screening and selection. *Annu. Rev. Biochem.* **2010**, *79*, 563−590.

(6) Ding, K.; Chin, M.; Zhao, Y.; et al. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nat. Commun.* **2024**, *15*, No. 6392.

(7) EUR-Lex - 52024DC0137 - EN - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52024DC0137.

(8) Frazier, P. I. A tutorial on Bayesian optimization. arXiv:1807.02811. arXiv.org e-Print archive. https://arxiv.org/abs/1807.02811. 2018.

(9) Gietz, R. D.; Schiestl, R. H.; Willems, A. R.; et al. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. *Yeast* **1995**, *11*, 355−360.

(10) Gustavsson, M.; Lee, S. Y. Prospects of microbial cell factories developed through systems metabolic engineering. *Microb. Biotechnol.* **2016**, *9*, 610−617.

(11) HamediRad, M.; Chao, R.; Weisberg, S.; et al. Towards a fully automated algorithm driven platform for biosystems design. *Nat. Commun.* **2019**, *10*, No. 5150.

(12) Head, T.; MechCoder, G. L.; Shcherbatyi, I.et al. scikit-optimize/scikit-optimize: v0. 5.2 *Version v0* 2018; Vol. *5*.

(13) Jeong, C.; Han, S. H.; Lim, C. G.; et al. Metabolic engineering of *Escherichia coli* for enhanced production of *p*-coumaric acid via L-phenylalanine biosynthesis pathway. *Bioprocess Biosyst. Eng.* **2025**, *48*, 565−576.

(14) Jeschek, M.; Gerngross, D.; Panke, S. Combinatorial pathway optimization for streamlined metabolic engineering. *Curr. Opin. Biotechnol.* **2017**, *47*, 142−151.

(15) Jeschek, M.; Gerngross, D.; Panke, S. Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat. Commun.* **2016**, *7*, No. 11163.

(16) Kang, C. K.; Shin, J.; Cha, Y.; et al. Machine learning-guided prediction of potential engineering targets for microbial production of lycopene. *Bioresour. Technol.* **2023**, *369*, No. 128455.

(17) Kaur, J.; Kaur, R. *p*-Coumaric acid: A naturally occurring chemical with potential therapeutic applications. *Curr. Org. Chem.* **2022**, *26*, 1333−1349.

(18) Kim, G. B.; Choi, S. Y.; Cho, I. J.; et al. Metabolic engineering for sustainability and health. *Trends Biotechnol.* **2023**, *41*, 425−451.

(19) Konzock, O.; Nielsen, J. TRYing to evaluate production costs in microbial biotechnology. *Trends Biotechnol.* **2024**, *42* (11), 1339−1347.

(20) Koren, S.; Walenz, B. P.; Berlin, K.; et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722−736.

(21) Lao-Martil, D.; Schmitz, J. P.; Teusink, B.; et al. Elucidating yeast glycolytic dynamics at steady state growth and glucose pulses through kinetic metabolic modeling. *Metab. Eng.* **2023**, *77*, 128−142.

(22) Lee, S. Y.; Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat. Biotechnol.* **2015**, *33*, 1061−1072.

(23) van Lent, P.; Schmitz, J.; Abeel, T. Simulated design-build-test-learn cycles for consistent comparison of machine learning methods in metabolic engineering. *ACS Synth. Biol.* **2023**, *12*, 2588−2599.

(24) Lewis, N. E.; Hixson, K. K.; Conrad, T. M.; et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **2010**, *6*, 390.

(25) Li, F.-Z.; et al. Evaluation of machine learning-assisted directed evolution across diverse combinatorial landscapes. *Cell Syst.* **2024**, *16* (9), No. 101387.

(26) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2021**, *23*, 18.

(27) Liu, Q.; et al. Modular pathway rewiring of yeast for amino acid production. *Methods Enzymol.* **2018**, *608*, 417−439.

(28) Liu, Q.; Yu, T.; Li, X.; et al. Rewiring carbon metabolism in yeast for high level production of aromatic chemicals. *Nat. Commun.* **2019**, *10*, No. 4976.

(29) Lu, H.; Li, F.; Sánchez, B. J.; et al. A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **2019**, *10*, No. 3586.

(30) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions *Adv. Neural Inf. Process. Syst.* 2017; Vol. *30* DOI: 10.48550/arXiv.1705.07874.

(31) Maftahi, M.; et al. Sequencing analysis of a 24.7 kb fragment of yeast chromosome XIV identifies six known genes, a new member of the hexose transporter family and ten new open reading frames. *Yeast* **1995**, *11.11*, 1077−1085.

(32) Mao, J.; Mohedano, M. T.; Fu, J.; et al. Fine-tuning of *p*-coumaric acid synthesis to increase (2*S*)-naringenin production in yeast. *Metab. Eng.* **2023**, *79*, 192−202.
(33) Moreno-Paz, S.; van der Hoek, R.; Eliana, E.; et al. Combinatorial optimization of pathway, process and media for the production of *p*-coumaric acid by *Saccharomyces cerevisiae*. *Microb. Biotechnol.* **2024**, *17*, No. e14424.
(34) Moreno-Paz, S.; van der Hoek, R.; Eliana, E.; et al. Machine learning-guided optimization of *p*-coumaric acid production in yeast. *ACS Synth. Biol.* **2024**, *13*, 1312−1322.
(35) Nitzan, S.; Rubinstein, A. A further characterization of Borda ranking method. *Public Choice* **1981**, *36*, 153−158.
(36) Noor, E.; Haraldsdóttir, H. S.; Milo, R.; et al. Consistent estimation of Gibbs energy using component contributions. *PLoS Comput. Biol.* **2013**, *9*, No. e1003098.
(37) Opgenorth, P.; Costello, Z.; Okada, T.; et al. Lessons from two design-build-test-learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. *ACS Synth. Biol.* **2019**, *8*, 1337−1351.
(38) Pandi, A.; Diehl, C.; Yazdizadeh Kharrazi, A.; et al. A versatile active learning workflow for optimization of genetic and metabolic networks. *Nat. Commun.* **2022**, *13*, No. 3876.
(39) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(40) Radivojević, T.; Costello, Z.; Workman, K.; et al. A machine learning Automated Recommendation Tool for synthetic biology. *Nat. Commun.* **2020**, *11*, No. 4879.
(41) Ren, J.; Zhang, M.; Yu, C.et al.Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2022; pp 7926−7935.
(42) Rodriguez, A.; Kildegaard, K. R.; Li, M.; et al. Establishment of a yeast platform strain for production of *p*-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis. *Metab. Eng.* **2015**, *31*, 181−188.
(43) Roubos, J. A.; Van Peij, N. N. M. E. Method for achieving improved polypeptide expression. US Patent US8,812,247, 2014.
(44) Sabzevari, M.; Szedmak, S.; Penttilä, M.; et al. Strain design optimization using reinforcement learning. *PLoS Comput. Biol.* **2022**, *18*, No. e1010177.
(45) Schomburg, I.; Jeske, L.; Ulbrich, M.; et al. The BRENDA enzyme information system-From a database to an expert system. *J. Biotechnol.* **2017**, *261*, 194−206.
(46) Steininger, M.; Kobs, K.; Davidson, P.; et al. Density-based weighting for imbalanced regression. *Mach. Learn.* **2021**, *110*, 2187−2211.
(47) Sutton, R. S. Reinforcement learning: An introduction. In *A Bradford Book* 2018.
(48) Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **2023**, *51*, D638−D646.
(49) Vaser, R.; Sović, I.; Nagarajan, N.; et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737−746.
(50) Verwaal, R.; Buiting-Wiessenhaan, N.; Dalhuijsen, S.; et al. CRISPR/Cpf1 enables fast and simple genome editing of *Saccharomyces cerevisiae*. *Yeast* **2018**, *35*, 201−211.
(51) Wick, R.; Volkening, J. Porechop: adapter trimmer for Oxford Nanopore reads. 2018, Available online: github com/rrwick/Porechop.
(52) Wright, C.; Wykes, M. Medaka: sequence correction provided by ONT research. 2023.
(53) Young, E. M.; Zhao, Z.; Gielesen, B. E.; et al. Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast. *Metab. Eng.* **2018**, *48*, 33−43.
(54) Zhang, J.; Petersen, S. D.; Radivojevic, T.; et al. Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* **2020**, *11*, No. 4880.

https://doi.org/10.1021/acssynbio.5c00864
ACS Synth. Biol. 2026, 15, 3146−3157