











# Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism

Jie Zhang <sup>1,13</sup>, Søren D. Petersen <sup>1,13</sup>, Tijana Radivojevic <sup>2,3,4</sup>, Andrés Ramirez <sup>5</sup>, Andrés Pérez-Manríquez <sup>5</sup>, Eduardo Abeliuk<sup>6</sup>, Benjamín J. Sánchez<sup>1</sup>, Zak Costello<sup>2,3,4</sup>, Yu Chen <sup>7,8</sup>, Michael J. Fero <sup>6</sup>, Hector Garcia Martin <sup>2,3,4,9</sup>, Jens Nielsen<sup>1,7,10</sup>, Jay D. Keasling <sup>1,2,3,11,12</sup> & Michael K. Jensen <sup>1</sup>✉

Through advanced mechanistic modeling and the generation of large high-quality datasets, machine learning is becoming an integral part of understanding and engineering living systems. Here we show that mechanistic and machine learning models can be combined to enable accurate genotype-to-phenotype predictions. We use a genome-scale model to pinpoint engineering targets, efficient library construction of metabolic pathway designs, and high-throughput biosensor-enabled screening for training diverse machine learning algorithms. From a single data-generation cycle, this enables successful forward engineering of complex aromatic amino acid metabolism in yeast, with the best machine learning-guided design recommendations improving tryptophan titer and productivity by up to 74 and 43%, respectively, compared to the best designs used for algorithm training. Thus, this study highlights the power of combining mechanistic and machine learning models to effectively direct metabolic engineering efforts.

<sup>1</sup>Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs., Lyngby, Denmark. <sup>2</sup>Joint BioEnergy Institute, Emeryville, CA, USA. <sup>3</sup>Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. <sup>4</sup>DOE Agile BioFoundry, Emeryville, CA, USA. <sup>5</sup>TeselaGen SpA, Santiago, Chile. <sup>6</sup>TeselaGen Biotechnology, San Francisco, CA, USA. <sup>7</sup>Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden. <sup>8</sup>Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden. <sup>9</sup>BCAM, Basque Center for Applied Mathematics, Bilbao, Spain. <sup>10</sup>BioInnovation Institute, Ole Maaløes Vej 3, DK-2200 Copenhagen N, Denmark. <sup>11</sup>Department of Chemical and Biomolecular Engineering & Department of Bioengineering, University of California, Berkeley, CA, USA. <sup>12</sup>Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen, China. <sup>13</sup>These authors contributed equally: Jie Zhang, Søren D. Petersen. ✉email: [mije@biosustain.dtu.dk](mailto:mije@biosustain.dtu.dk)

Metabolic engineering is the directed improvement of cell properties through the modification of specific biochemical reactions<sup>1</sup>. Beyond offering an improved understanding of basic cellular metabolism, the field of metabolic engineering also envisions sustainable production of biomolecules for health, food, and manufacturing industries, by fermenting feedstocks into value-added biomolecules using engineered cells<sup>2</sup>. These promises leverage tools and technologies developed over recent decades that include both nonintuitive evolution-guided approaches, such as adaptive laboratory evolution<sup>3,4</sup>, as well as rational approaches combining mechanistic metabolic modeling, targeted genome engineering, and robust bioprocess optimization; ultimately aiming for accurate and scalable predictions of cellular phenotypes from deduced genotypes<sup>5</sup>.

Among the different types of mechanistic models for simulating metabolism, genome-scale models (GSMs) are one of the most popular approaches, as they are genome complete, covering thousands of metabolic reactions. These computational models not only provide qualitative mapping of cellular metabolism<sup>6,7</sup>, but have also been successfully applied for the discovery of metabolic functions<sup>8</sup>, and to guide engineering designs toward desired phenotypes<sup>9</sup>. As GSMs are built based only on the stoichiometry of metabolic reactions, several methods have been developed to account for additional layers of information, regarding the chemical intermediates and the catalyzing enzymes participating in the metabolic pathways of interest<sup>10</sup>. Nevertheless, all these mechanistic models require a priori knowledge, as well as high-quality data for accurate prediction<sup>11,12</sup>.

Machine learning (ML) provides a complementary approach to guide metabolic engineering by learning patterns on system behavior from large experimental datasets<sup>13</sup>. As such, ML models differ from mechanistic models by being purely data-driven. Indeed, ML methods for the generation of predictive models on living systems are becoming ubiquitous, including applications within genome annotation, de novo pathway discovery, product maximization in engineered microbial cells, pathway dynamics, and transcriptional drivers of disease states<sup>14</sup>. While being able to provide predictive power based on complex multivariate relationships<sup>15</sup>, the training of ML algorithms requires large datasets of high quality, and thereby imposes certain standards for the experimental workflows. For instance, for genotype-to-phenotype predictions, it is desirable that datasets contain a high variation between both genotypes and phenotypes<sup>16</sup>. Also, measurements on the individual experimental unit, e.g., a strain, should be accurate and obtainable in a high-throughput manner, in order to limit the number of iterative design–build–test–learn cycles needed to reach the desired output.

While mechanistic models require a priori knowledge of the living system of interest, and ML-guided predictions require ample multivariate experimental data for training, the combination of mechanistic and ML models holds promise for improved performance of predictive engineering of cells by uniting the advantages of the causal understanding of mechanism from mechanistic models, with the predictive power of ML<sup>15,17</sup>. Metabolic pathways are known to be regulated at multiple levels, including transcriptional, translational, and allosteric levels<sup>13</sup>. To cost-effectively move through the design and build steps of complex metabolic pathways, combinatorial optimization of metabolic pathways, in contrast to sequential genetic edits, has been demonstrated to effectively facilitate the searching for global optima for outputs of interest (i.e., production<sup>18</sup>). Searching global optima using combinatorial approaches involves facing an exponentially growing number of designs (known as the combinatorial explosion) and requires efficient building of multi-parameterized combinatorial libraries. However, this challenge can be mitigated by using intelligently designed condensed

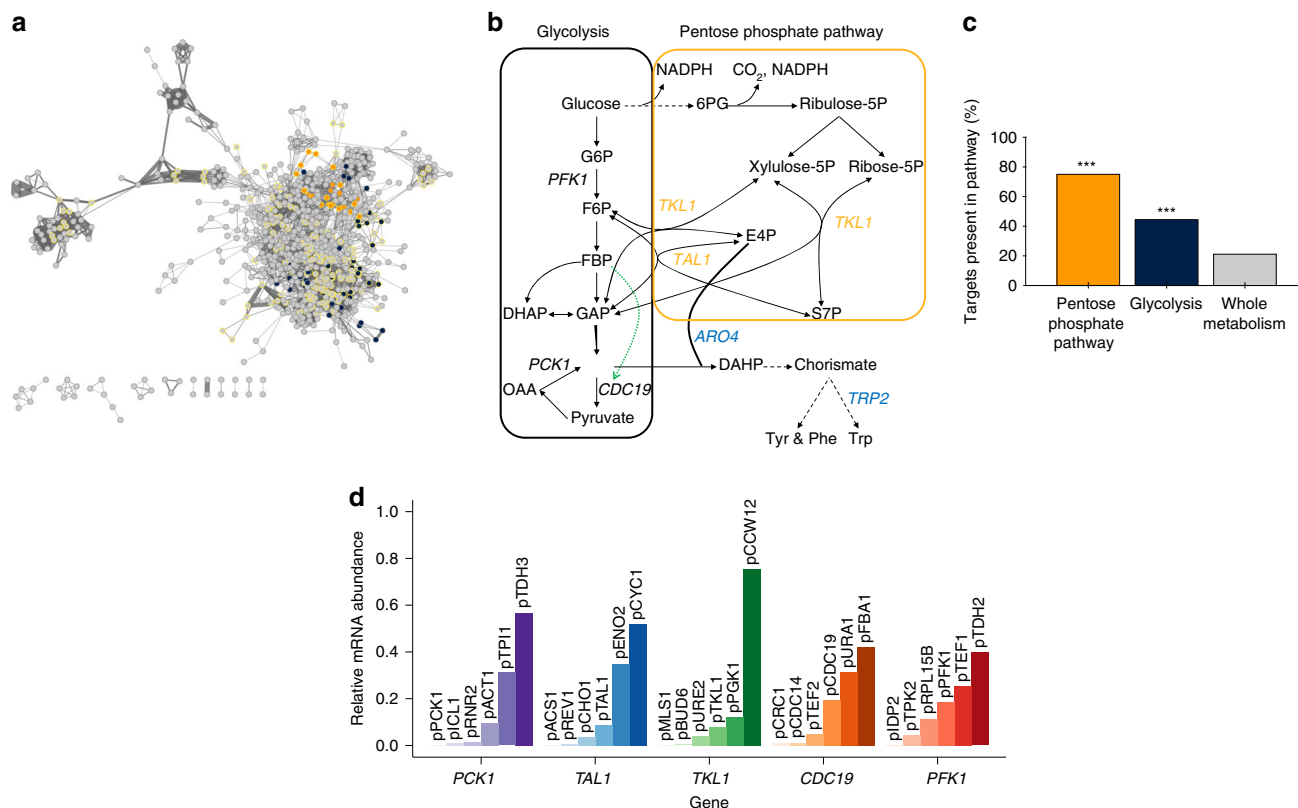
libraries that allow uniform discretization of multidimensional spaces: e.g., by using well-characterized sets of DNA elements controlling the expression of candidate genes at defined levels as opposed to using more less-/non-characterized random elements<sup>19,20</sup>. As cellular metabolism is regulated at multiple levels<sup>21,22</sup>, an efficient search strategy for global optima using combinatorial approaches should also take this into consideration, e.g., by using mechanistic models, “omics data repositories”, and a priori biological understanding. Still it should be noted, that even with intelligent choice of design parameters and efficient library construction, there is no guarantee mathematical models will reach such a global optimum.

Here we combine mechanistic and ML models to enable robust genotype-to-phenotype predictions as a tool for metabolic engineering. The approach is exemplified for predictive engineering and optimization of the complexly regulated aromatic amino acid (AAA) pathway that produces tryptophan in baker’s yeast *Saccharomyces cerevisiae*<sup>23</sup>. We define a 7776-membered combinatorial library design space, based on five genes selected from GSM simulations and a priori biological understanding, each controlled by six different promoters from a set of 30 promoters selected from transcriptomics data mining. To train predictive models for tryptophan biosynthesis rate in yeast, we collect >124,000 experimental time series data points derived from fluorescent read-outs of an engineered tryptophan biosensor encoded into >500 different strain designs. This enable selection of optimal sampling time points, from that we explore fluorescence synthesis rates of ~3% (250/7,776) of the possible genetic designs of the library design space. Based on genotype data, growth profiles, and the biosensor output, we train various ML algorithms. Predictive models based on these algorithms identify designs exhibiting up to 74% higher tryptophan titers than best designs used for training the models.

## Results

**Model-guided design of high tryptophan production.** One prime example of the multitiered complexity regulating metabolic fluxes is the shikimate pathway, driving the central metabolic route leading to AAA biosynthesis<sup>24</sup>. This pathway has enormous industrial relevance, since it has been used to produce bio-based replacements of a wealth of fossil fuel-derived aromatics, polymers, and potent human therapeutics<sup>25</sup>.

To search for gene targets to perturb tryptophan production, we initially performed constraint-based modeling for predicting single gene targets, with a simulated objective of combining growth and tryptophan production<sup>26</sup>. From this analysis, we retrieved 192 genes, covering 259 biochemical reactions, which showed considerable changes as production shifted from growth toward tryptophan production (Fig. 1a, b, Supplementary Data 1). By performing an analysis for statistical overrepresentation of genome-scale modeled metabolic pathways, we observed that both the pentose phosphate pathway (PPP) and glycolysis were among the top pathways with a significantly higher number of gene targets compared to the representation of all metabolic genes (Fig. 1c, Supplementary Data 2). Among the predicted gene targets in those pathways, *CDC19*, *TKL1*, *TAL1*, and *PCK1* were initially selected as targets for combinatorial library construction (Fig. 1b), as these genes have all been experimentally validated to be directly linked or to have an indirect impact on the shikimate pathway precursors erythrose 4-phosphate (E4P) and phosphoenolpyruvate (PEP). Specifically, *CDC19* encodes the major isoform of pyruvate kinase converting PEP into pyruvate to fuel the tricarboxylic acid cycle, while *TKL1* and *TAL1* that encode the major isoform of transketolase and transaldolase, respectively, in the reversible non-oxidative PPP, have been reported to impact



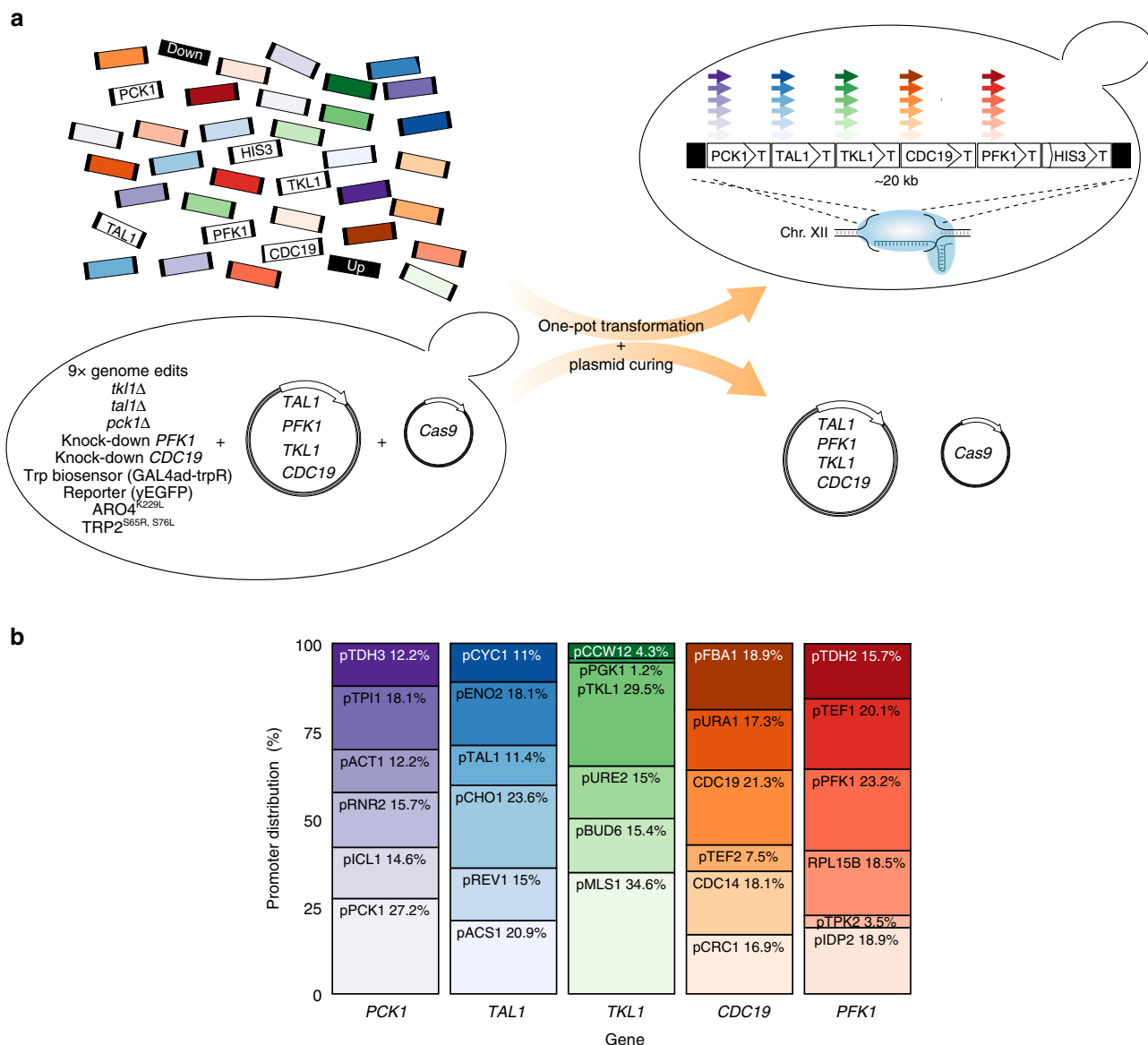
**Fig. 1 Gene targets and promoters for combinatorial engineering of tryptophan metabolism in *S. cerevisiae*.** **a** Gene-gene interaction network built with Cytoscape, showing that pentose phosphate pathway and glycolysis are both in the core of metabolism in close proximity to many genes. Nodes are all 909 genes in yeast metabolism<sup>67</sup>, sharing connections based on the number of shared metabolites by the corresponding reactions that the genes are related to: the thicker the edge, the higher the number of shared metabolites. Currency metabolites such as water, protons, ATP, etc. are removed from the analysis. The prefuse force directed layout is used for displaying the network. Genes are highlighted with a yellow border if they are selected targets by the mechanistic modeling approach, and in orange and dark blue if they belong to the pentose phosphate pathway or glycolysis, respectively. **b** Simplified map of metabolism showing the selected gene targets from glycolysis (dark blue) and pentose phosphate pathway (orange) based on a combination of mechanistic genome-scale modeling and literature studies for optimizing tryptophan production. Black dashed lines indicate multistep reactions. Dashed green line indicates allosteric activation. G6P glucose 6-phosphate, F6P fructose 6-phosphate, FBP fructose 1,6-bisphosphate, GAP glyceraldehyde 3-phosphate, DHAP dihydroxyacetone phosphate, PEP phosphoenolpyruvate, OAA oxaloacetate, 6PG 6-phosphogluconate, E4P erythrose 4-phosphate, S7P sedoheptulose 7-phosphate, DAHP 3-deoxy-7-phosphoheptulonate, Tyr tyrosine, Phe phenylalanine, Trp tryptophan. **c** Percentage of genes in glycolysis (dark blue) and pentose phosphate pathway (orange) that were predicted by the mechanistic modeling to increase tryptophan production compared to the percentage of genes predicted as targets from the whole metabolism. \*\*\*P-value < 0.05, two-sided Fisher’s exact testing with  $n = 54$  and  $24$  for the glycolysis and pentose phosphate pathway, respectively. **d** Relative messenger RNA (mRNA) abundance, calculated for each gene as the proportion of mRNA reads obtained for any given promoter relative to the total sum of mRNA reads from each bin of six promoters. Absolute abundances for the 30 promoters were measured in *S. cerevisiae* CEN.PK113-7D in the mid-log phase<sup>32</sup>. The promoters are grouped according to intended combinatorial gene associations. Source data underlying Fig. 1d are provided as a Source data file.

the supply of E4P<sup>27,28</sup>. In addition, focusing on the E4P and PEP linkage, *PCK1* encoding PEP carboxykinase, was also selected due to its regeneration capacity of PEP from oxaloacetate<sup>29</sup>. Lastly, while not being predicted as a target by the constraint-based modeling approach, the *PFK1* gene, encoding the alpha subunit of heterooctameric phosphofructokinase (PFK1), catalyzing the irreversible conversion of fructose 6-phosphate to fructose 1,6-bisphosphate (FBP), was selected, as insufficient activity of this enzyme is known to cause divergence of carbon flux toward the PPP across different kingdoms<sup>30,31</sup>.

Next, we mined transcriptomics datasets for the selection of promoters to control the expression of the five target genes. Here, we focused on well-characterized and sequence-diverse promoters to ensure rational designs spanning large absolute levels of promoter activities, and limit the risk of recombination within strain designs and loss of any genetic elements, respectively<sup>32,33</sup> (Supplementary Fig. 1). Together, this mining resulted in the selection of 25 sequence-diverse promoters, which together with

the five promoters natively regulating the target genes constitutes the parts catalog for combinatorial library design (Fig. 1d, Supplementary Fig. 1, Supplementary Table 1).

**Creation of a platform strain for a combinatorial library.** To construct a combinatorial library targeting equal representation of 30 promoters expressing five target genes, we harnessed high-fidelity homologous recombination in yeast together with the targetability of CRISPR/Cas9 genome engineering for a one-pot assembly of a maximum of 7776 ( $6^5$ ) different combinatorial designs. Due to the dramatic decrease in transformation efficiency when simultaneously targeting multiple loci in the genome<sup>34</sup>, we targeted the sequential deletion of all five selected target genes from their original genomic loci, and next assembled a cluster of five expression cassettes into a single genomic landing as recently successfully reported for the single-locus glycolysis in yeast<sup>35</sup> (Fig. 2a; see “Methods” section). However, as *CDC19* is an



**Fig. 2 Construction and validation of the 13-parts assembled 20 kb combinatorial promoter:gene library. a** Strategy for library construction including a 13-part in vivo assembly for the reintegration of target genes into a single genomic locus. The platform strain used for one-pot transformation includes a total of nine genome edits for knockout, knockdown, and heterologous expression of candidate genes (see “Methods” section). yEGFP yeast-enhanced green fluorescent protein. **b** Promoter distribution (name, % representation) by gene. Color intensity correlates with promoter strength (see Fig. 1d). Source data underlying Fig. 2b are provided as a Source data file.

**Table 1 Key descriptive statistics for the library construction and genotyping.**

Potential unique genotypes	7776
Number of library colonies	-10,000
Number of colonies sampled	480
Cured strains (%)	92
Correct assembly (%)	82
Repeated genotypes (%)	3.7

essential gene, and deletion of *PFK1* causes growth retardation<sup>36,37</sup>, our platform strain for library construction had a galactose-curable plasmid introduced expressing *PFK1*, *CDC19*, *TKL1*, and *TAL1* under their native promoters (see “Methods” section), after deleting *PCK1*, *TKL1*, and *TAL1*, and knocking down *CDC19* and *PFK1* (Fig. 2a). Prior to one-pot assembly of

the combinatorial library, we integrated the two feedback-resistant shikimate pathway enzymes 3-deoxy-D-arabinose-heptulosonate-7-phosphate (DAHPS) synthase (*ARO4*<sup>K229L</sup>) and anthranilate synthase (*TRP2*<sup>S65R, S76L</sup>) into the platform strain<sup>38,39</sup>, known to increase AAA accumulation in microbial cells<sup>40,41</sup>.

**One-pot construction of the combinatorial library.** For library construction, we transformed in one-pot the platform strain with 38 different parts (30 promoters, 5 ORFs, *HIS3* ORF, and 2 homology regions) for 7776 unique 20 kb 13-parts assemblies at the targeted genomic locus (Fig. 2a). To assess assembly fidelity and ensure benchmarking, we also transformed yeast with five user-defined clusters, including one design with native promoters in front of each of the five selected genes (herein labeled the reference strain; Supplementary Table 2). Following transformation, we randomly sampled 480 colonies from the library,



together with 27 colonies from the five control strains (507 in total), and successfully cured 423 out of 461 (92%) sufficiently growing strains of the complementation plasmid by means of galactose-induced expression of the dosage-sensitive gene *ACT1* (ref. 42; Fig. 2b, Supplementary Fig. 6). Next, genotyping identified 380 out of 461 (82%) of the sufficiently growing strains to be correctly assembled with only 9 out of 245 (3.7%) of the fully filtered library genotypes observed in duplicates (245 = 250 library and control genotypes—five control genotypes; Table 1, Supplementary Fig. 2). Based on a Monte Carlo simulation with 10,000 repeated samplings of 10,000 library colonies, and assuming percent correct assemblies and promoter distribution as determined for the library sample (Fig. 2), the expected number of unique genotypes among all library colonies was calculated to be 3759, equaling estimated library coverage of 48% (3759/7776). Importantly, all 30 promoters from the one-pot transformation were represented in the genotyped designs, with promoters *PGK1* (no. 14) and *MLS1* (no. 15), represented the least (1%) and most (35%), respectively (Fig. 2b).

Taken together, these results demonstrate high transformation efficiency of the platform strain, high fidelity of parts assembly, and expected high coverage of the genetically diverse combinatorial library design.

### A biosensor for high-throughput library characterization.

In order to support high-throughput analysis of tryptophan accumulation in library strains, we harnessed the power of modular engineering allosterically regulated transcription factors as small-molecule biosensors<sup>43</sup>. Here, a yeast tryptophan biosensor was developed based on the *trpR* repressor of the *trp* operon from *Escherichia coli*<sup>44</sup>. We first tested *trpR*-mediated transcriptional repression by expressing *trpR* together with a GFP reporter under the control of the strong *TEF1* promoter, containing a palindromic consensus *trpO* sequence<sup>45</sup> (5'-GTACTAGTT-AAC-TAGTAC-3') downstream of the TATA-like element<sup>46</sup> (TATTTAAG; Fig. 3a). From this, we observed that *trpR* was able to repress GFP expression by 2.4-fold (Supplementary Fig. 3a). Next, to turn the native *trpR* repressor into an activator with positively correlated biosensor-tryptophan readout, we fused the Gal4 activation domain to the N-terminus of *trpR* (*GAL4<sub>AD</sub>-trpR*) under the control of the weak *REVI* promoter (Supplementary Fig. 3). For the reporter promoter, we placed *trpO* 97 bp upstream of the TATA-like element of the *TEF1* promoter (Supplementary Fig. 3b) and observed that *trpR* was able to activate GFP expression by a maximum of 1.75-fold upon supplementing tryptophan to the cultivation medium (Supplementary Fig. 3b). To further optimize the dynamic range of the reporter output, the GFP reporter was expressed under a hybrid promoter consisting of tandem repeats of triple *trpO* sequences (i.e., in total 6 × *trpO* sequences) located 88 bp upstream of the TATA box in an engineered *GAL1* core promoter without Gal4 binding sites, ultimately enabling *GAL4<sub>AD</sub>-trpR*-mediated biosensing with a dynamic output range of fivefold, and an operational input range spanning supplemented tryptophan concentrations from ~2–200 mg/L (Fig. 3b).

To further validate the designed biosensor, we measured fluorescence output in strains engineered for expression of feedback-resistant versions of *ARO4* and *TRP2* (refs. 38,39; *ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R,S76L</sup>), and observed high biosensor outputs from these strains in line with previously demonstrated high enzyme activities in strains expressing *ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R,S76L</sup> (refs. 38,39), and thus corroborating the ability of the tryptophan biosensor to monitor changes in endogenously produced tryptophan pools (Fig. 3c). Most importantly, we confirmed the biosensor readout as a valid proxy for tryptophan

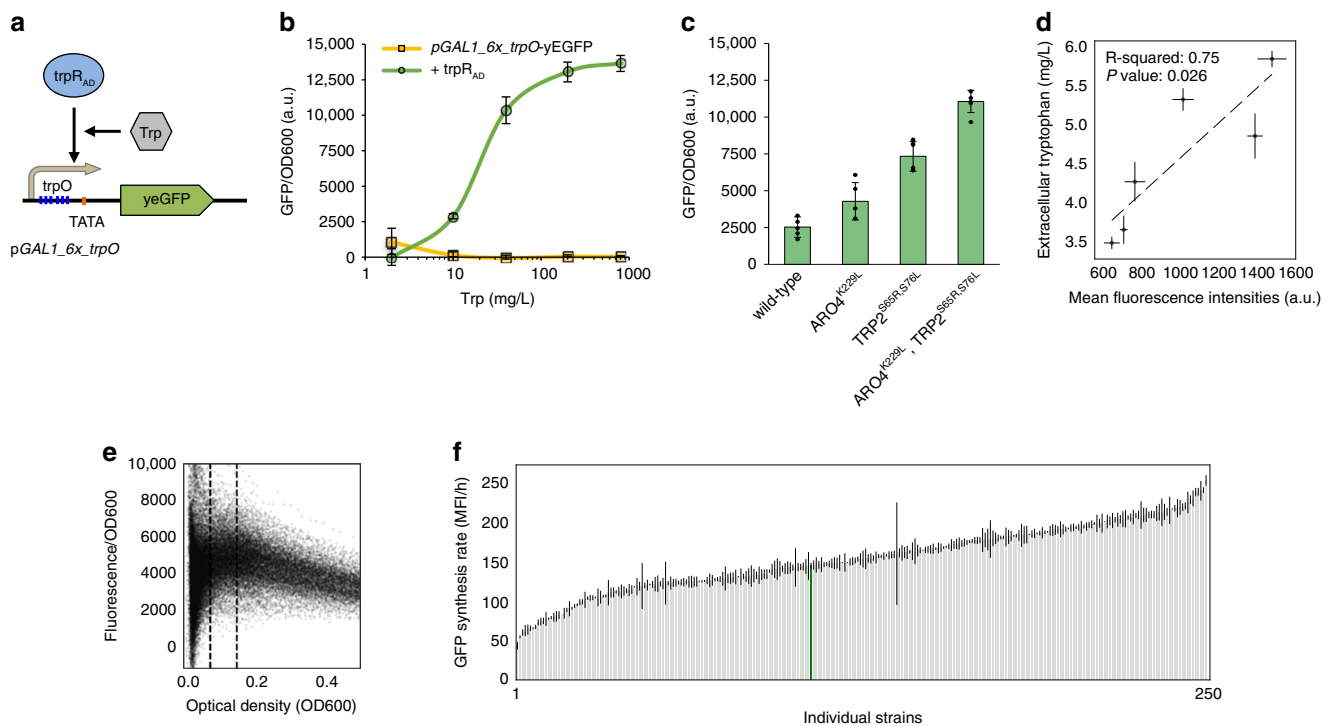
levels, by comparing external tryptophan titers measured by HPLC with a change in GFP intensities for six library strains spanning 2.5-fold changes in GFP intensities ( $R^2 = 0.75$ ; Fig. 3d).

Having established a biosensor for high-throughput screening of the combinatorial library, we next sought to explore the maximal resolution of the biosensor readout at the single-design level of growing isoclonal strains, with the intention to define optimal data sampling time point. To do so, we measured time series data of OD and GFP at 82 time points in triplicates for all 507 colonies (that is 480 from the library and 27 from the control strains), covering a total of 124,722 data points (Supplementary Figs. 4 and 5). Here, as we observed that the fluorescence per cell generally stabilized at an OD value of 0.075 and started to decrease beyond an OD value of 0.15 (Fig. 3e, Supplementary Fig. 4, see “Methods” section), and the between strains variation in fluorescence at the single-cell level was relatively high within this OD interval, we chose this interval for determining the GFP synthesis rate as a proxy for tryptophan biosynthesis rate. The average GFP synthesis rate of all quality-controlled strains (see below) was observed to vary between 43.7 and 255.7 MFI/h (approximately sixfold; Fig. 3f), with an average standard error of the mean of 6.6 MFI/h corresponding to an average coefficient of variation for the mean values of 4.3%. By comparison, the GFP synthesis rate of the platform strain, expressing *ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R,S76L</sup> together with all five candidate genes under native promoters, was 144.8 MFI/h (Fig. 3f).

### Using machine learning to predict metabolic pathway designs.

Having successfully established a combinatorial genetic library and a large phenotypic dataset thereof, we next assessed the potential of using ML to predict promoter combinations expected to improve tryptophan productivity. Since there is no single algorithm that is optimal for all conceivable general learning tasks<sup>47</sup>, we decided to improve our chances by using two different ML approaches for the single regression learning task of predicting promoter combinations controlling five genes that best improve GFP biosynthesis rates, as a proxy for tryptophan productivity: the Automated Recommendation Tool (ART) and EVOLVE algorithm<sup>48,49</sup> (see “Methods” section). Briefly, ART uses a Bayesian ensemble approach where eight regressors from the scikit-learn library<sup>50</sup> are allowed to vote on a prediction with a weight proportional to their accuracy; the EVOLVE algorithm is inspired by Bayesian Optimization and uses an ensemble of estimators as a surrogate model that predicts the outcome of the process to be optimized (see “Methods” section). As the quality of the data is of paramount importance for ML predictions, data were initially filtered in order to avoid strains (i) with insufficient growth, (ii) without sequencing data, (iii) with incorrect assembly, (iv) without plasmid curation, or (v) that exhibited more than one genotype (see “Methods” section; Supplementary Fig. 5). Following this, ~58% (266/461) of the growing strains remained after filtering, while another 3% of the remaining data were removed because of lack of reproducibility (high error in triplicate measurements), ultimately leaving high-quality sequencing and GFP data from 250 genotypes as input training dataset (Supplementary Fig. 5).

Both modeling approaches, ART and EVOLVE, were able to recapitulate the data they were trained on. The average (obtained from ten independent runs) training mean absolute error (MAE) of the predicted tryptophan production compared to the measured values was 13.8 and 11.9 MFI/h for the ART and EVOLVE model approaches, respectively, when calculated for the whole dataset (Fig. 4a, b). These MAEs represent ~7 and 6% of the full range of measurements (50–200 MFI/h). The train MAE uncertainty (represented by the shaded area in Fig. 4a, b and



**Fig. 3 Phenotypic library characterization using an engineered tryptophan biosensor.** **a** Schematic illustration of the design of the tryptophan (Trp) biosensor ( $\text{trpR}_{\text{AD}}$ ) engineered in this study. The  $\text{trpR}_{\text{AD}}$  indicates the engineering tryptophan biosensor composed of the *E. coli* TrpR fused to the GAL4 activation domain. The biosensor regulates an engineered reporter (yeGFP) *GAL1* promoter, including 6 $\times$  copies of TrpR binding sites (*trpO*), placed upstream of the TATA box of *GAL1* promoter (*pGAL1\_6x\_trpO*). **b** Fluorescence normalized by optical density ( $\text{OD}_{600}$ ) for two strains related to concentration of tryptophan supplemented media (mean fluorescence intensity/ $\text{OD}$ , MFI/ $\text{OD}$  with standard errors,  $n = 3$  biological replicates). Both strains contain the yeGFP reporter under the control of the *pGAL1\_6x\_trpO* reporter promoter, and only one strain expresses the Gal4 activation domain fused to trpR (in green). **c** Fluorescence normalized by  $\text{OD}_{600}$  for a wild-type strain and strains with expression of feedback-resistant versions of *ARO4* and *TRP2*, *ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R,S76L</sup>, respectively (mean fluorescence intensity, MFI/h with standard errors,  $n = 4$ –5 biological replicates). **d** Extracellular tryptophan normalized by  $\text{OD}_{600}$  related to fluorescence normalized by  $\text{OD}_{600}$  (mean values with standard errors,  $n = 3$  technical replicates). The  $p$ -value showing a significant slope is from a two-sided  $t$ -test performed on mean values for the six different genotypes. **e** Fluorescence divided by  $\text{OD}_{600}$  related to  $\text{OD}_{600}$  for library and control strains. Dashed lines are shown at  $\text{OD}_{600}$  equals 0.075 and 0.15. **f** Measured mean green fluorescent protein synthesis rate. MFI/h with standard errors,  $n = 3$  technical replicates. The data is ranked according to increasing mean rate. The strain with five native promoters expressing the five candidate genes is highlighted in green. GFP green fluorescent protein, MFI mean fluorescence intensity,  $\text{OD}_{600}$  optical density (600 nm), a.u. arbitrary units. Source data underlying Fig. 3b–f are provided as a Source data file.

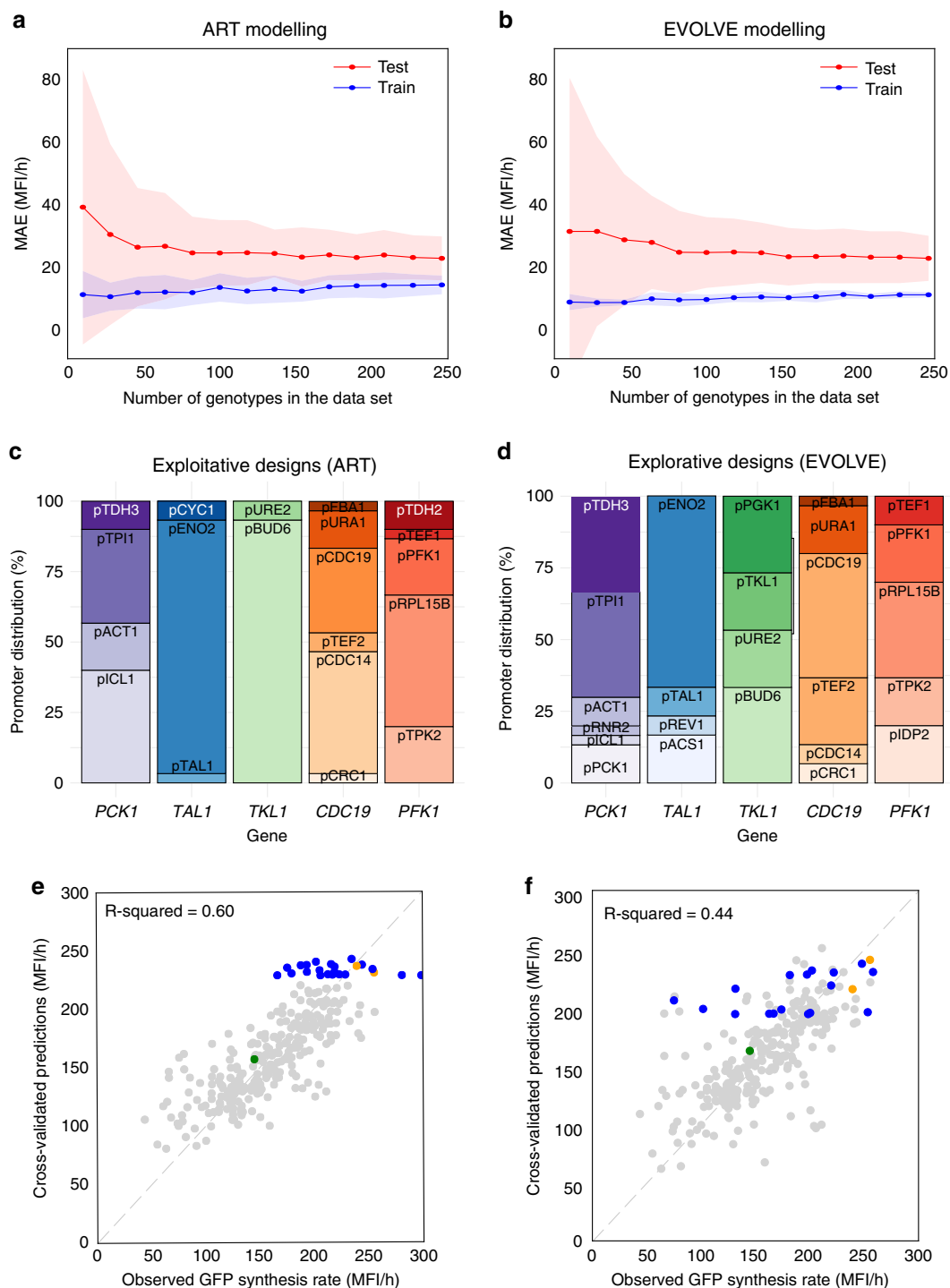
quantified as the 95% confidence interval from ten runs) decreased slightly with increasing size of the training dataset for ART, whereas the overall uncertainty was smaller for the EVOLVE model approach (Fig. 4a, b). The ability to predict the production for new promoter combinations the algorithms had not been trained on was tested by cross-validation, i.e., by training the model on 90% of the data, and then testing the predictions of this model against measurements for the remaining 10% (tenfold cross-validation). Here, the average cross-validated MAE (test MAE) was 21.4 and 22.4 MFI/h for ART and EVOLVE model approaches, respectively (Fig. 4a, b), which represent ~11% of the full range of measurements. The test MAE decreased systematically with the size of the dataset, yet the decrease rate declined markedly as more data was added. However, while the two approaches had similar average cross-validated MAEs, the uncertainty of the MAEs was slightly smaller for ART than for the EVOLVE algorithm (Fig. 4a, b).

**Predictive engineering of high tryptophan production.** Next, beyond enabling prediction of tryptophan production, we used an exploitative approach implemented in the ART model and an explorative one adopting the EVOLVE algorithm to recommend two sets of 30 prioritized designs aiming for high tryptophan production (Supplementary Tables 3 and 4). The exploitative model

focuses on exploiting the predictive power to recommend promoter combinations that improve production, whereas the explorative model combines predictive power with the estimated uncertainty of each prediction, to recommend promoter combinations<sup>48,49</sup>.

Among the recommendations from each of the two ML approaches, two overlapped (SP588 and SP627, Supplementary Tables 3 and 4). Interestingly, while use of *PGK1* promoter to control *TKL1* expression was underrepresented in the original library sample (Fig. 2b), the explorative set of recommendations included eight (even top three) designs with *PGK1* promoter for expression control of *TKL1*, and the exploitative approach included none (Supplementary Data 2; Fig. 4c, d). From construction of these recommendations, we used the same genome engineering approach as for library construction (Fig. 2a) to successfully construct 19 individual assemblies of the explorative recommendations and 24 individual assemblies of the exploitative recommendations. Interestingly, we were not able to construct any of the eight designs with the *PGK1* promoter, partially explaining the lower number of viable strains found with the explorative approach.

Of the 41 recommendations constructed, the predictions from both sets generally fitted well with the measurements, and both approaches successfully enabled predictive strain engineering for high-performing GFP synthesis rates, with the best recommendation (SP606) having a measured GFP synthesis rate 106% higher



**Fig. 4 Machine learning-guided predictive engineering of tryptophan metabolism.** **a, b** Learning curves for ART and EVOLVE algorithms, respectively. Mean absolute error (MAE) from model training and testing as a function of the number of genotypes in the dataset. Shaded areas represent 95% confidence intervals based on ten random samples of the given no of genotypes ( $n$ ). Blue curves indicate MAE when calculated for the whole dataset (train), while red curves indicate the cross-validation, i.e., by training the models on 80% of the data and then testing the predictions of this model against measurements for the remaining 20% (test). **c, d** Promoter distributions for the 30 recommendations of the exploitative (ART) and explorative (EVOLVE) approach, respectively. The orders and colors of promoters correspond to those in Fig. 1c. **e, f** Cross-validated predictions vs average of measured GFP synthesis rate for the exploitative (ART) and explorative (EVOLVE) approach, respectively. Data are shown for library and control strains (gray markers; green markers show the platform strain expressing *ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R,S76L</sup>), as well as for recommended strains (blue markers; orange markers show recommendations that overlap between the two approaches).  $R$ -squared values are for cross-validated predictions for the whole dataset (not only training set data). MFI mean fluorescence intensity. Source data are provided as a Source data file.

than the already improved platform design (SP507), and 17% higher than the best one (SP271) in the library sample (Fig. 4e, f). This has been confirmed by HPLC analysis from small-scale deep-well batch cultivations of a diverse set of control, library, and recommended strains (Supplementary Fig. 7). We observed the strain SP606 having a 74 and 43% improvement in tryptophan titer and productivity, respectively, compared to the best strain design from the library sample (SP271). Moreover, eight recommendations were found in the top ten of productivity, of which four were from the exploitative set, three were from the explorative set, and one overlapping between the two sets. Comparing the output of the ART and EVOLVE approaches, the variation in measurements was higher for strains recommended with the explorative EVOLVE approach than for strains recommended with the exploitative ART approach (Fig. 4e, f), and the explorative approach included recommendations based on a more diverse set of promoters than the exploitative approach (Fig. 4c, d). Aligned with this, we observed that the recommendations from the EVOLVE approach also included a fraction of combinatorial designs with GFP synthesis rates below the reference strain (Fig. 4f). Still, taken together, when run in parallel, ART and EVOLVE approaches successfully enable predictive engineering of tryptophan biosynthesis strain designs, and for both approaches even strains with tryptophan biosynthesis rates beyond those previously observed for training the models (Fig. 4e, f, Supplementary Table 5 and 6).

## Discussion

In this study, we focus on the current possibility of using mechanistic and ML-guided models for predictive engineering of cellular metabolism as compared to sequential trial-and-error metabolic engineering iterations, or adaptive evolution-based reverse engineering for identification of nonintuitive changes. From this, we demonstrated that mechanistic and ML approaches can complement and enhance each other, enabling a more effective predictive engineering of living systems. Using a single design-build-test-learn cycle, this study (i) leveraged mechanistic GSMs to select and rank reactions/genes most likely to affect production, (ii) included the efficient one-pot construction of a library with different promoter combinations controlling the expression of these genes, and (iii) used ML algorithms trained on the ensuing phenotyping data to choose promoter combinations that further enhance tryptophan productivity. In total, we managed to increase tryptophan titers and productivity by up to 74% and 43%, respectively, compared to an already improved reference strain (*ARO4*<sup>K229L</sup> and *TRP2*<sup>S65R, S76L</sup>).

To gather the large high-quality dataset required for ML approaches, we developed a biosensor that enabled the sampling of >124,000 GFP intensity measurements (82 time points) as a proxy for tryptophan flux for 1521 isoclonal designs (three replicates × 507 strains) in a high-throughput fashion, of which data from 250 strains were eventually used for successful training of ML algorithms (Fig. 3e, Supplementary Fig. 5a). Indeed, while requiring a few design iterations (Fig. 3a, Supplementary Fig. 3), the tryptophan biosensor ultimately allowed us to (i) phenotypically characterize an order of magnitude higher number of strains than in previous ML-guided metabolic engineering studies<sup>20,51–53</sup>, and (ii) identify optimal sampling points that displayed the largest differences between genotypes (Fig. 3c, Supplementary Fig. 4). Likewise, one-pot CRISPR/Cas9-mediated genome editing was a vital enabling technology for this project, since it allowed us to efficiently create a diverse 20-kb clustered combinatorial library with representation of all 30 specified sequence- and expression-diverse promoters to control five expression units, including very few duplicate designs (Fig. 2b, Table 1).

Enabled by this high-quality dataset, we used two different ML models for predicting productivity (ART and EVOLVE algorithm), and two different approaches to recommend strain designs (exploitative and explorative). Cross-validation showed that both models could be trained to show good correlations (MAE ~ 11% of the measurement range) between predictions and measurements for data they had not seen previously (test data). The test MAE decreased considerably with the number of genotypes in the dataset, this decrease was similar for both models. With this in mind, a relevant guideline for choosing a recommendation approach should focus on the desired outcome: the explorative approach providing a more diverse set of recommendations (Fig. 4c, d), whereas the exploitative approach provides less varied recommendations. We observed the largest improvement in titer and productivity when using the exploitative approach (Fig. 4e, f, Supplementary Fig. 7). However, if subsequent design-build-test-learn cycles are performed, the diversity of recommendations of the explorative approach could help avoid local optima of tryptophan production (Fig. 4e, f).

Notably, while the recommendations were able to improve biosynthesis rates, the predictions from both ML models were noticeably worse than for the library, reflecting the general challenge of extrapolating outside of the previous range of measurements. As such, we envision that future ML approaches will need to focus on models able to extrapolate more efficiently.

Another critical aspect to discuss from this study is the amount and quality of data needed, in order to increase the impact (e.g., improving titers, rates, and yields) and reduce model uncertainty. From this study, we argue that biosensors for time-resolved sensing of cellular metabolism not only enable sampling of large amounts of data points, but most importantly also facilitate the identification of a smaller sampling space for high-quality determination of metabolite biosynthesis rates (Fig. 3e). Specifically, we initially sampled triplicate measurements for 82 time points for all 576 strains, which when compared to growth, ultimately allowed us to select 15 time points of relevance for calculating maximal GFP biosynthesis rates. Likewise, while the one-pot library construction used in this study had an estimated coverage of 48% of the full combinatorial design space, the amount of strains used for training the algorithms only covered ~3%, yet enabled predictive engineering following a single design-build-test-learn cycle. This could be used to argue that more engineering iterations on even smaller datasets, potentially coupled to mixed exploitation and exploration approaches as recently demonstrated for cell-free production<sup>54</sup>, should be a valid avenue for ML-guided engineering of even less genetically tractable chassis, and for which no high-throughput screening method may even exist. With regards to this, we performed a follow-up test running the ART and EVOLVE approaches in explorative and exploitative modes, respectively. Here, we observed that the recommendations from EVOLVE in exploitative mode had overlaps of 20% (6/30) and 23% (7/30) to ART recommendations in exploitative, and EVOLVE recommendations in explorative mode, respectively. Complementary to this, the recommendations from ART in explorative mode only had overlaps of 3% (1/30) and 0% (0/30) to ART recommendations in exploitative and EVOLVE recommendations in explorative mode, respectively (Supplementary Fig. 8), indicating that the uncertainty of prediction of high GFP synthesis rate weighted differently for the two models in explorative mode.

While discovery of strain designs with titers and rates out-competing previously reported high aromatics producers was not the main motivation for the study, it should be mentioned that all strains tested in this study produce much lower mg/L levels of tryptophan compared to previous studies, focusing on metabolic engineering and bioprocess optimization for aromatics



overproduction (Fig. 3d, Supplementary Fig. 7)<sup>23</sup>. Indeed, as a suggestion for further optimization, it is possible that the reference strain used in this study is still subject to certain levels of feedback inhibition, as suggested by recent studies for AAAs derivatives<sup>25,55</sup>. Furthermore, the use of fed-batch cultivations as part of a bioprocess optimization would also be expected to enable cells to accumulate higher tryptophan titers compared to the titers obtained based on short batch cultivations in 96-well deep plates with low oxygen levels used in this study.

Despite the low production, there is still a positive correlation between tryptophan titer/productivity and the GFP synthesis rate (Fig. 3c, d, Supplementary Fig. 7), and the large-scale dataset from this study provides examples of results anticipated based on rational engineering, as well as nonintuitive predictions, enabling further advancement of the biological understanding of tryptophan metabolism. For instance, the best-performing strain (SP606, Supplementary Table 3 and Supplementary Data 3) predicted by ML, included knockdowns of both *CDC19* and *PFK1*, corroborating our intuitive strategies for increasing precursor availability: i.e., lower pyruvate kinase activity would lead to higher PEP pools, while limiting glycolysis redirects carbon flux into PPP and subsequently increases E4P<sup>27</sup>. Indeed, in bacteria, pyruvate kinase knockout has been used for the overproduction of shikimate pathway-derived aromatics products in bacteria<sup>56–58</sup>. Likewise, since yeast cells with *CDC19* deletion cannot grow on glucose<sup>59</sup>, dynamic silencing of *CDC19* and *PYK2* have been used for boosting production of para-hydroxybenzoic acid<sup>60</sup>, just as expression of a mutant *CDC19* pyruvate kinase with seemingly lower activity, in combination with overexpression of transketolase (*TKL1*), have been demonstrated to improve 2-phenylethanol production in yeast<sup>61</sup>. On the contrary, a similar strategy with lower *CDC19* activity, but in combination with *zwf1Δ* deletion (lacking the committed step toward the oxidative branch of PPP) was shown to reduce tyrosine titers<sup>62</sup>. Surprisingly, the top five strains predicted to have high tryptophan biosynthesis rates (SP606, SP616, SP624, SP588, and SPSP620, Supplementary Tables 3 and 4), all had low expression of *TKL1* and high expression of *TAL1*, despite the report that overexpression of *TKL1*, rather than *TAL1*, leads to higher AAA production in both *E. coli* and yeast<sup>27,61</sup>. These discrepancies remark the importance of carefully considering the systems-level context of these metabolic rules-of-thumb (e.g., overexpress *TKL1* instead of *TAL1* for higher amino acid production) to ensure their validity. Consistently, both the second (SP616) and third (SP624) best-performing strains, also predicted by ML, had low expression of *TKL1* and high expression of *TAL1*, together with very low expression (*TPK2* promoter) for *PFK1* and high expression of *CDC19*. One possible explanation is that, although normally expressed, the pyruvate kinase activity could be limited by the low level of its allosteric activator FBP due to the limited PFK expression. Another plausible explanation is that medium-high expression of *PCK1* (conversion of oxaloacetate to PEP) by *ACT1* or *TDH3* promoters in these two strains can replenish PEP pools consumed by pyruvate kinase. The fact that eight out of ten top-performing strains had high expression of *PCK1* (Supplementary Data 3), which was not predicted to be impactful on glucose by the GSM approach, indicates that this indeed has a positive effect on tryptophan biosynthesis rate, and stresses the importance of combining mechanistic and ML approaches.

Ultimately, in our case study, ML models have demonstrated good performance in predicting GFP biosynthesis rates for the training data designs (gray dots in Fig. 4e, f), while the recommended strains' biosynthesis rates were less accurately predicted, likely because it involved an extrapolation effort that is a known weakness for ML methods (blue dots in Fig. 4e, f). In spite of this decrease in predictive power, the ML models can effectively

recommend designs that improve tryptophan biosynthesis rates (Supplementary Fig. 7). However, this predictive power is heavily dependent on the availability of high-quality experimental data, which is not a prerequisite for mechanistic GSMs. Without any experimental input, GSMs are able to guide metabolic engineering using various constraint-based algorithms, which, however, predict a large number of potential targets and may also miss some effective ones (e.g., *PFK1* in our study), due to the lack of other information beyond metabolism, e.g., regulation in GSMs. To address this problem, manual efforts are currently needed to filter out less relevant targets and add intuitively promising ones based on existing knowledge. In addition, applying our approach to new models that enhance GSMs with more levels of information, such as kinetics<sup>63</sup>, gene expression<sup>64</sup>, and regulation<sup>65</sup> is envisioned to further improve the model's predictive power.

Irrespective of the ongoing efforts for model-guided engineering of living cells, this study highlights the enhanced predictive power from combining GSMs for selecting genetic targets with ML algorithms for leveraging experimental data. Finally, as even more efficient methods for combining data-driven ML algorithms and GSMs are developed, we envision accelerated improvements in our ability to engineer virtually any cell system effectively.

## Methods

**Experimental models.** *S. cerevisiae* strains were derived from CEN.PK2-1C (EUROSCARF, Germany). These were cultivated in yeast synthetic dropout media (Sigma-Aldrich) at 30 °C. *E. coli* DH5α were cultivated in LB medium containing 100 mg/L ampicillin (Sigma-Aldrich) at 37 °C.

**Mechanistic modeling of high tryptophan flux.** In order to select targets for increased tryptophan accumulation, we followed a constraint-based strategy implemented in a recent study<sup>66</sup>. Briefly, flux balance analysis (FBA)<sup>26</sup> was used to simulate growth of *S. cerevisiae* at 11 different suboptimal growth conditions ranging from 30 to 80% of the maximum specific growth rate, with all remaining flux oriented toward tryptophan accumulation. Based on these simulations, a score was calculated for each reaction in metabolism as the average simulated flux fold change compared to maximum growth rate conditions. These reaction scores were in turn used to compute gene scores, by averaging the associated reaction scores. A gene score higher than one means that the gene is associated with reactions that increase in flux as tryptophan production increases and could point to a target for overexpression. On the other hand, a gene score lower than one signifies that the gene is connected to reactions that decrease their flux as tryptophan production increases, and therefore could be a target for downregulation. The analysis was performed with either glucose or ethanol as carbon sources, so to find candidates under a mixed-fermentation regime, a purely respiratory regime and the overlap between both regimes. The seventh version of the consensus GSM of *S. cerevisiae*<sup>67</sup>, a parsimonious FBA approach<sup>68</sup>, and the COBRA toolbox<sup>69</sup> v. 3.0.6 were used for all simulations.

**Promoter selection.** Each of the five gene targets was expressed under six unique promoters. The six promoters included the promoter native to the gene, as well as five promoters chosen to span a wide expression range. All promoters were chosen based on absolute mRNA abundances measured for *S. cerevisiae* CEN.PK113-7D in the mid-log phase<sup>32</sup>, and unless otherwise stated were 1 kb in length by default. To minimize homologous recombination during one-pot transformation for library construction and potential loop out of promoters and genes following genomic integration, all scanned promoter sequences were aligned to ensure there were no extensive homologous sequence stretches.

**General strain construction.** Strains were edited using the CasEMBLR method<sup>70</sup>. All integrations were directed toward EasyClone sites<sup>71</sup>. Homology regions between DNA parts were by default 30 bp, and homology regions, framing the repair assembly, were ~0.5 kb. Yeast transformations were performed by LiAc/SS carrier DNA/PEG method<sup>70</sup>. DNA parts and plasmids were purified using kits from Macherey-Nagel. PCR products for USER assembly were amplified using Phusion U Hot Start PCR Master Mix (ThermoFisher), bricks for transformation by Phusion High-Fidelity PCR Master Mix with HF Buffer (ThermoFisher), whereas colony PCRs were performed using 2× OneTaq Quick-Load Master Mix with Standard Buffer (New England Biolabs). Genomic DNA was extracted from overnight cultures using Yeast DNA Extraction Kit (Thermo Scientific). Oligos were purchased from IDT. Sequencing was performed by Eurofins. All primers, plasmids, and yeast strains, are listed in Supplementary Data 4 and Supplementary Tables 7 and 8, respectively.

**Platform strain construction.** As *CDC19* is an essential gene, and deletion of *PFK1* causes growth retardation<sup>36,37</sup>, this genetic background was deemed unsuitable for efficient one-pot transformation. For this reason our platform strain for library construction had a galactose-curable plasmid introduced expressing *PFK1*, *CDC19*, *TKL1*, and *TAL1* under their native promoters, before performing two sequential rounds of CRISPR-mediated genome engineering to delete *PCK1*, *TKL1*, and *TAL1*, and knockdown *CDC19* and *PFK1* using the weak promoters *RNR2* and *REV1*, respectively (Fig. 2a). Moreover, several enzymes within the AAA biosynthesis are subject to allosteric regulations. Specifically, DAHP synthase (encoded by *ARO4*), which controls the entry of the shikimate pathway, is feedback inhibited by all three AAAs, although to different extents<sup>72,73</sup>. Anthranilate synthase (encoded by *TRP2*), which catalyzes the first committed step toward the tryptophan branch, is also inhibited by its end product tryptophan<sup>24</sup>. To maximize the transcriptional regulatory effect on the tryptophan flux, and benchmark with current state-of-the-art in shikimate pathway optimization, feedback-resistant variants of these two enzymes, *ARO4*<sup>K229L38</sup> and *TRP2*<sup>S65R, S76L39</sup>, were overexpressed under the *TEF1* and *TDH3* promoters, respectively, at EasyClone site XI-3 (ref. 71) (Supplementary Table 8). Lastly, a tryptophan biosensor system was introduced by integrating corresponding sensor and reporter sequences into EasyClone sites at Chr. XI-2 and XI-5, respectively<sup>71</sup>.

**Construction of combinatorial library.** Due to the dramatic decrease in transformation efficiency targeting multiple loci in the genome<sup>34</sup>, we opted for removing all five target genes from their original loci and assemble the five expression units into a single cluster for targeted integration into EasyClone site XII-5 (ref. 71), and thereby ensuring comparable genomic accessibility of all genes. While *PCK1*, *TKL1*, and *TAL1* were successfully knocked out, deleting *PFK1* and/or *CDC19* was unsuccessful. Alternatively, we replaced *PFK1* and *CDC19* promoters with weak *REV1* and *RNR2* promoters, respectively. Due to an expected loss of activity in *PFK1* and pyruvate kinase (*CDC19*), and consequently slow ATP generation, the resulting strain (TrpNA-W) grew extremely poorly and was barely transformable using linear DNA fragments for assembly. To overcome this limitation, the TrpNA-W strain was complemented with plasmid pCfB9307 (Supplementary Table 8) harboring *PFK1*, *CDC19*, *TKL1*, and *TAL1* genes, which restored the growth to the wild-type level. The plasmid backbone carries yeast *ACT1* gene under the control of *GAL1* promoter, which can be used as counterselection of the plasmid due to the growth arrest caused by *ACT1* overexpression on galactose as the sole carbon source<sup>42</sup> (Supplementary Fig. 6).

For combinatorial library construction, we adopted CasEMBLR<sup>70</sup>. Briefly, five target genes together with a *HIS3* expression cassette (in the order of *PCK1-TAL1-TKL1-CDC19-PFK1-HIS3*) were assembled in the same orientation and integrated at EasyClone site XII-5<sup>71</sup>. All five target genes (the complete ORFs) together with their terminators (500 bp downstream of the stop codon) were amplified from the genomic DNA of yeast strain CEN.PK113-7D using primers listed in Supplementary Data 4. All 30 promoters (defined as the 1000 bp upstream of the ORF) were amplified using primers with a 30 bp overlap to adjacent DNA parts (i.e., the terminator upstream and the target gene). All promoters can be found in Supplementary Table 1. The *HIS3* cassette was amplified from plasmid pRS413-*HIS3* (ref. 71) with primers 30 bp overlapping with the *PFK1* terminator and fragment homologous to the downstream of XII-5. The *HIS3* cassette was included as one part of the assembly. The one-pot transformation of all 38 parts (30 promoters, 5 candidate genes, *HIS3* cassette, and up- and down-homology regions for EasyClone site XII-5) was performed with 50 mL the base strain grown to an optical density of 1.0 (equivalent to 6.5 mg of cell dry weight), 5.0 µg of plasmid expressing the guide RNA targeting XII-5, and 1.0 picomole of each of 13 DNA fragments. A total of 480 colonies were picked from ten transformation plates by dividing the area of each individual plate into four subareas of equal size and picking 12 colonies of varying size from each subarea.

Finally, the complementation plasmid introduced was cured by culturing strains to stationary phase twice in media with galactose instead of glucose as carbon source (Supplementary Fig. 6). The success of curing was then gauged by a growth assay where LEU auxotrophs were considered as cured and prototrophs as not cured. Control strains and recommended strains were constructed similarly to the library strains except that instead of transforming pools of promoter parts for each gene only specific promoters were transformed per gene.

**Development of tryptophan biosensor.** The yeast tryptophan biosensor was developed based on the *trpR* repressor of the *trp* operon from *E. coli*<sup>44</sup>. The *trpR* gene was amplified from *E. coli* M1665 genome. All yeast promoters as well as the activator domain of *GAL4* were amplified from *S. cerevisiae* strain CEN.PK113-7D genome. All designs of *trpR* biosensor and GFP reporter were first cloned into the pRS416 (*URA3*) and pRS413 (*HIS3*) vectors, respectively, by USER cloning (NEB). The activator domain of *GAL4* (*GAL4<sub>AD</sub>*) was fused to *trpR* with a GSGSGS linker by USER cloning, and the expression of the gene product controlled by the weak *REV1* promoter. The *trpO* sequence was inserted into the *TEF1* promoter 8 bp downstream of the TATA-like element (TATTAAAG) by inverse PCR from a plasmid containing the *P<sub>TEF1</sub>-yEGFP-T<sub>ADH1</sub>* cassette, with both primers containing the overhang AACTAGTAC (i.e., half of the *trpO* sequence). The linear PCR product was treated with DpnI enzyme to fragment the template plasmid and self-ligated to generate circular plasmid (Quick Ligation™ Kit, NEB). Promoters

containing multiple *trpO* sequences were constructed by USER cloning from a synthetic DNA fragment (Integrated DNA Technologies) of a minimal *GAL1* promoter (−329 to −5 relative to the *GAL1* open reading frame, thus without the *GAL4* binding sequence that is located at −435 to −418) with 3× tandem repeats of *trpO* (separated by two nucleotides) inserted at 88 bp upstream of the TATA box (TATATAAA). Plasmids containing the sensor and reporter cassettes were transformed into yeast strain CEN.PK113-11C. To test the biosensor performance, yeast transformants were grown in selection media overnight and regrown in Delft medium supplemented with various tryptophan concentrations (2–1000 mg/L) for 6 h (typically reaching early exponential phase). GFP and mKate2 outputs were measured on Synergy MX microtiter plate reader (BioTek) with excitation/emission at 485/515 nm and 588/633 nm, respectively, and always normalized by absorbance at 600 nm (OD<sub>600</sub>). To construct the base strain for library assembly, the tryptophan sensor (*P<sub>REV1</sub>-GAL4<sub>AD</sub>-trpR-T<sub>ADH1</sub>*) and the reporter cassette (*P<sub>GAL1core\_3xtrpO</sub>-yEGFP-T<sub>ADH1</sub>*, *P<sub>TEF1</sub>-trpO-mKate2-T<sub>CYC1</sub>*) were integrated into strain TC-3 (ref. 34) at the EasyClone sites XI-2 and XI-5<sup>71</sup>, respectively.

**Validation of biosensor by HPLC.** To validate the correlation between biosensor reporter gene output and tryptophan production, we quantified extracellular tryptophan concentrations by HPLC<sup>74</sup>. Supernatants of cultivated strains were separated from the culture broth using AcroPrep Advance 96-Well Filter Plates (Pall Corporation) and centrifugation (5 min at 2200 × g) following 24 h of cultivation in synthetic dropout medium without tryptophan and histidine. From this, 200 µl was used for analysis on a Dionex 3000 HPLC system with a Zorbax Eclipse Plus C18 column (Agilent Technologies, Santa Clara, CA, USA). The column temperature was set to 30 °C. The flow rate was set to 1 ml/min with a mobile phase consisting of 0.05% acetate and a variable amount of acetonitrile. The total duration per sample was 12 min. This consisted of 10 min of separation, in which acetonitrile was reduced from 95 to 38.7% in 9.4 min and held for 0.6 min, 1 min of returning the acetonitrile concentration to 95%, and 1 min of holding the concentration till the end of the run. The injection volume was set to 10 µl. Elution of tryptophan was detected by UV at a wavelength of 280 nm. The data were processed using Chromleon™ Chromatography Data System Software v7.1.3. Tryptophan concentrations were determined from a calibration curve. The specific tryptophan productivity was estimated as the average amount of tryptophan secreted into the medium per unit of biomass during the period of 1 day (µmol/gDCW/day).

**DNA sequencing of assembled clusters.** Genomic DNA was extracted from overnight cultures using the LiOAc/SDS method adapted to a 96-well microtiter plate format. Each extract was used as a template in five PCR reactions spanning the five integrated promoters and amplifying from 1200 to 1700 bp. The PCR products were validated using a LabChip GX II (Perkin Elmer) and sequenced using PlateSeq PCR Kits (Eurofins) according to the manufacturer's instructions. From the LabChip results, a PCR reaction was considered as trusted if it showed a strong band of the correct size; not trusted if it showed a strong band of the wrong size, and as no information (NI) gained if it showed a weak or no band. From the sequencing results, a sequencing reaction was considered as trusted if it showed an unambiguous sequence of the expected length (i.e., only limited by length of PCR fragment, stretches of the same nucleotide in the promoter or of ~1 kb limit of sanger sequencing reactions), not trusted if it showed an unambiguous sequence of the expected length with an assembly error, and NI gained if there were no or bad sequence results. If one or more sequencing results from the same strain showed double peaks in the promoter region the strain was considered as a double population. Finally, the promoter was noted as a failed assembly if either LabChip and/or sequencing results were considered not trusted, as NI if the sequencing result was NI and else as the promoter predicted by pairwise alignment between sequencing results and promoter sequence.

**Measuring fluorescence and growth.** Yeast cells were cultured O/N to saturation, diluted to OD<sub>600</sub> 0.025 (measured by reading the absorbance at 600 nm on Synergy Mx Microplate Reader, BioTek) and then cultured again in a Synergy Mx Microplate Reader. While culturing, the reader measured OD<sub>600</sub> and fluorescence with excitation and emission wavelengths of 485 and 515 nm, respectively, every 15 min for 20 h. All wells were sealed with VIEWseal membrane (Greiner Bio-One).

**Modeling and recommendation.** All genotype and time series data as well as scripts for preprocessing are publicly available (see “Data and Code availability” sections). Briefly, all OD<sub>600</sub> and GFP measurements were subtracted background signals (i.e., mean value of OD<sub>600</sub> and GFP measurements in wells containing pure media). Background signals were calculated for each 96-well plate. Strains were quality-controlled based on five criteria. The criteria were: (1) optical densities must cover the whole range up to 0.15 OD<sub>600</sub> units to exclude uninoculated wells and wells with insufficient growth, (2) sequencing results must exist for all five promoter gene junctions, (3) the integrated sequence must be exactly as designed, (4) the complementation plasmid must be cured, and (5) the sequencing results must not indicate the presence of multiple genotypes (Supplementary Fig. 5a). Specific GFP synthesis rates were calculated as the difference in GFP divided by the difference in time (MFI/h) in the OD<sub>600</sub> interval from 0.075 to 0.150, as measured by a Synergy Mx Microplate

Reader from BioTek (a detailed description of the rationale behind this method can be found in connection with Supplementary Fig. 4).

In the ART approach, outliers were identified and removed based on replicate differences in GFP synthesis rate relative to the mean value for the strain. Replicates with the one percent most extreme differences were identified and the corresponding strains were removed. GFP synthesis rate was modeled as a function of promoter combination, represented through one-hot encoding, using the ART<sup>48</sup>. Briefly, ART uses a probabilistic ensemble model consisting of eight individual models. The weight of each ensemble model is considered a random variable with a probability distribution characterized by the available training data, and determined through Bayesian inference and Markov Chain Monte Carlo<sup>75</sup>. ART uses the trained ensemble model in combination with a Parallel Tempering approach<sup>76</sup> to recommend 30 promoter combinations (unseen designs), which are predicted to improve production. The recommended designs were chosen as the 30 strains with the highest expected GFP synthesis rate predicted by the model. This recommendation approach was labeled exploitative since predictions with high uncertainty were not prioritized, although ART can provide both exploitative and explorative recommendations.

For the TeselaGen EVOLVE algorithm used in this study, outliers were identified and removed based on a method described by Rousseuw and Hubert<sup>77</sup>. The decision was made on a per strain basis taking into account replicate to mean value differences. In cases where just a single replicate was left after filtering, this replicate was excluded as well. Of the remaining strains, GFP synthesis rates were modeled as a function of promoter combination coded as categorical variables using a TeselaGen-developed ML algorithm based on Bayesian Optimization<sup>78</sup>. The algorithm was set up to recommend 30 promoter combinations (unseen designs), and designs were chosen by highest selection score. The selection score was the expected improvement<sup>79</sup>, calculated based on predicted high GFP synthesis rate and the uncertainty of prediction. The approach was labeled explorative since high uncertainty weighed positively in the selection score calculation. While using EVOLVE for explorative recommendations, thereby complementing the ART approach, it should be mentioned that EVOLVE can be set up to provide both explorative and exploitative recommendations.

For both approaches, we tried encoding the promoter variables both as numbers ordered according to the counts from the RNAseq experiment (i.e., promoter strength<sup>32</sup>) and as one-hot encoding, and chose the one-hot encoding because it produced a lower MAE values and higher *R*-squared values.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this article is available as a Supplementary Information file. The datasets generated and analyzed during the current study are available from the corresponding author upon request. The genotype and time series datasets are available at The Joint BioEnergy Institute's Inventory of Composable Elements (ICE; <https://public-registry.jbei.org>) and Experiment Data Depot (EDD; <https://public-edd.jbei.org>), respectively under the study "Zhang and Petersen, et al. 2019". These are also available at GitHub ([https://github.com/sorpet/Zhang\\_and\\_Petersen\\_et\\_al\\_2019](https://github.com/sorpet/Zhang_and_Petersen_et_al_2019)). Source data are provided with this paper.

## Code availability

The FBA, with additional simulation details and filtering criteria, is available at GitHub (<https://github.com/biosustain/trp-scores>). The preprocessing and statistical calculations were made in a Python v. 3.6.5 environment with the packages seaborn 0.7.1, scikit-learn 0.20.2, pymc3 3.5, pandas 0.23.4, numpy 1.14.3, matplotlib 3.0.2, scipy 1.1.0, PTMCMC Sampler 2015.2, and ART. The process is documented in a jupyter notebook, available at GitHub ([https://github.com/sorpet/Zhang\\_and\\_Petersen\\_et\\_al\\_2019](https://github.com/sorpet/Zhang_and_Petersen_et_al_2019)). The notebook also contains the ART approach for model development and strain recommendation. The Teselagen software is available through commercial and non-commercial licenses (<https://teselagen.com>).

Received: 30 December 2019; Accepted: 27 July 2020;

Published online: 25 September 2020

## References

- Stephanopoulos, G. Metabolic fluxes and metabolic engineering. *Metab. Eng.* **1**, 1–11 (1999).
- Keasling, J. D. Manufacturing molecules through metabolic engineering. *Science* **330**, 1355–1358 (2010).
- Reyes, L. H., Gomez, J. M. & Kao, K. C. Improving carotenoids production in yeast via adaptive laboratory evolution. *Metab. Eng.* **21**, 26–33 (2014).
- Sandberg, T. E., Salazar, M. J., Weng, L. L., Palsson, B. O. & Feist, A. M. The emergence of adaptive laboratory evolution as an efficient tool for biological discovery and industrial biotechnology. *Metab. Eng.* **56**, 1–16 (2019).
- Lee, J. W. et al. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* **8**, 536–546 (2012).
- Monk, J. M. et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* **35**, 904–908 (2017).
- Lu, H. et al. A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).
- Guzmán, G. I. et al. Model-driven discovery of underground metabolic functions in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **112**, 929–934 (2015).
- Yang, J. E. et al. One-step fermentative production of aromatic polyesters from glucose by metabolically engineered *Escherichia coli* strains. *Nat. Commun.* **9**, 79 (2018).
- Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305 (2012).
- Khodayari, A., Chowdhury, A. & Maranas, C. D. Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model. *Front. Bioeng. Biotechnol.* **2**, 76 (2015).
- Long, C. P. & Antoniewicz, M. R. Metabolic flux responses to deletion of 20 core enzymes reveal flexibility and limits of *E. coli* metabolism. *Metab. Eng.* **55**, 249–257 (2019).
- Chubukov, V., Gerosa, L., Kochanowski, K. & Sauer, U. Coordination of microbial metabolism. *Nat. Rev. Microbiol.* **12**, 327–340 (2014).
- Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C. & Collins, J. J. Next-generation machine learning for biological networks. *Cell* **173**, 1581–1592 (2018).
- Presnell, K. V. & Alper, H. S. Systems metabolic engineering meets machine learning: a new era for data-driven metabolic engineering. *Biotechnol. J.* **14**, 1800416 (2019).
- Carbonell, P., Radivojevic, T. & García Martín, H. Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth. Biol.* **8**, 1474–1477 (2019).
- Zampieri, G., Vijayakumar, S., Yaneske, E. & Angione, C. Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* **15**, e1007084 (2019).
- Jeschek, M., Gerngross, D. & Panke, S. Combinatorial pathway optimization for streamlined metabolic engineering. *Curr. Opin. Biotechnol.* **47**, 142–151 (2017).
- Jeschek, M., Gerngross, D. & Panke, S. Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat. Commun.* **7**, 11163 (2016).
- Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J. & Dueber, J. E. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.* **41**, 10668–10678 (2013).
- Feng, Y. et al. Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* **32**, 1036–1044 (2014).
- Lahtvee, P. J. et al. Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst.* **4**, 495–504.e5 (2017).
- Averesch, N. J. H. & Krömer, J. O. Metabolic engineering of the shikimate pathway for production of aromatics and derived compounds—present and future strain construction strategies. *Front. Bioeng. Biotechnol.* **6**, 32 (2018).
- Braus, G. H. Aromatic amino acid biosynthesis in the yeast *Saccharomyces cerevisiae*: a model system for the regulation of a eukaryotic biosynthetic pathway. *Microbiol. Rev.* **55**, 349–370 (1991).
- Liu, Q. et al. Rewiring carbon metabolism in yeast for high level production of aromatic chemicals. *Nat. Commun.* **10**, 4976 (2019).
- Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
- Curran, K. A., Leavitt, J. M., Karim, A. S. & Alper, H. S. Metabolic engineering of muconic acid production in *Saccharomyces cerevisiae*. *Metab. Eng.* **15**, 55–66 (2013).
- Patnaik, R. & Liao, J. C. Engineering of *Escherichia coli* central metabolism for aromatic metabolite production with near theoretical yield. *Appl. Environ. Microbiol.* **60**, 3903–3908 (1994).
- Yin, Z. Multiple signaling pathways trigger the exquisite sensitivity of yeast gluconeogenic mRNAs to glucose. *Mol. Microbiol.* **20**, 751–764 (1996).
- Wang, Y., San, K.-Y. & Bennett, G. N. Improvement of NADPH bioavailability in *Escherichia coli* through the use of phosphofructokinase deficient strains. *Appl. Microbiol. Biotechnol.* **97**, 6883–6893 (2013).
- Yi, W. et al. Phosphofructokinase 1 glycosylation regulates cell growth and metabolism. *Science* **337**, 975–980 (2012).
- Rajkumar, A. S. et al. Engineered reversal of function in glycolytic yeast promoters. *ACS Synth. Biol.* **8**, 1462–1468 (2019).



33. Reider Apel, A. et al. A Cas9-based toolkit to program gene expression in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **45**, 496–508 (2017).
34. Jakočiūnas, T. et al. Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. *Metab. Eng.* **28**, 213–222 (2015).
35. Kuijpers, N. G. A. et al. Pathway swapping: toward modular engineering of essential cellular processes. *Proc. Natl Acad. Sci. USA* **113**, 15060–15065 (2016).
36. Breslow, D. K. et al. A comprehensive strategy enabling high-resolution functional analysis of the yeast genome. *Nat. Methods* **5**, 711–718 (2008).
37. Cherry, J. M. et al. Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705 (2012).
38. Hartmann, M. et al. Evolution of feedback-inhibited / barrel isoenzymes by gene duplication and a single mutation. *Proc. Natl Acad. Sci. USA* **100**, 862–867 (2003).
39. Graf, R., Mehmman, B. & Braus, G. H. Analysis of feedback-resistant anthranilate synthases from *Saccharomyces cerevisiae*. *J. Bacteriol.* **175**, 1061–1068 (1993).
40. Park, S. H. et al. Metabolic engineering of *Corynebacterium glutamicum* for L-arginine production. *Nat. Commun.* **5**, 4618 (2014).
41. Vogt, M. et al. Pushing product formation to its limit: metabolic engineering of *Corynebacterium glutamicum* for L-leucine overproduction. *Metab. Eng.* **22**, 40–52 (2014).
42. Makanae, K., Kintaka, R., Makino, T., Kitano, H. & Moriya, H. Identification of dosage-sensitive genes in *Saccharomyces cerevisiae* using the genetic tug-of-war method. *Genome Res.* **23**, 300–311 (2013).
43. Rogers, J. K., Taylor, N. D. & Church, G. M. Biosensor-based engineering of biosynthetic pathways. *Curr. Opin. Biotechnol.* **42**, 84–91 (2016).
44. Gunsalus, R. P. & Yanofsky, C. Nucleotide sequence and expression of *Escherichia coli trpR*, the structural gene for the trp aporepressor. *Proc. Natl Acad. Sci. USA* **77**, 7117–7121 (1980).
45. Yang, J. et al. In vivo and in vitro studies of TrpR-DNA interactions. *J. Mol. Biol.* **258**, 37–52 (1996).
46. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
47. Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural Comput.* **8**, 1341–1390 (1996).
48. Radivojević, T., Costello, Z., Workman, K., & Martin, H. G. ART as a machine learning Automated Recommendation Tool for synthetic biology. *Nat. Commun.* <https://doi.org/10.1038/s41467-020-18008-4> (2020).
49. TeselaGen. *TeselaGen Technology Including EVOLVE Module* <https://teselagen.com> (2019).
50. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **6**, 2825–2830 (2011).
51. Alonso-Gutierrez, J. et al. Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng.* **28**, 123–133 (2015).
52. Redding-Johanson, A. M. et al. Targeted proteomics for metabolic pathway optimization: application to terpene production. *Metab. Eng.* **13**, 194–203 (2011).
53. Zhou, Y. et al. MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae*. *Metab. Eng.* **47**, 294–302 (2018).
54. Borkowski, O. et al. Large scale active-learning-guided exploration for *in vitro* protein production optimization. *Nat. Commun.* **11**, 1872 (2020).
55. Leavitt, J.M. et al. Biosensor-enabled directed evolution to improve muconic acid production in *Saccharomyces cerevisiae*. *Biotechnol. J.* **12**, 1600687 (2017).
56. Kitade, Y., Hashimoto, R., Suda, M., Hiraga, K. & Inui, M. Production of 4-hydroxybenzoic acid by an aerobic growth-arrested bioprocess using metabolically engineered *Corynebacterium glutamicum*. *Appl. Environ. Microbiol.* **84**, e02587-17(2018).
57. Licona-Cassani, C. et al. Inactivation of pyruvate kinase or the phosphoenolpyruvate: sugar phosphotransferase system increases shikimic and dehydroshikimic acid yields from glucose in *Bacillus subtilis*. *J. Mol. Microbiol. Biotechnol.* **24**, 37–45 (2014).
58. Meza, E., Becker, J., Bolivar, F., Gosset, G. & Wittmann, C. Consequences of phosphoenolpyruvate:sugar phosphotransferase system and pyruvate kinase isozymes inactivation in central carbon metabolism flux distribution in *Escherichia coli*. *Microb. Cell Factories* **11**, 127 (2012).
59. Sprague, G. F. Isolation and characterization of a *Saccharomyces cerevisiae* mutant deficient in pyruvate kinase activity. *J. Bacteriol.* **130**, 232–241 (1977).
60. Williams, T. C. et al. Quorum-sensing linked RNA interference for dynamic metabolic pathway control in *Saccharomyces cerevisiae*. *Metab. Eng.* **29**, 124–134 (2015).
61. Hassing, E.-J., de Groot, P. A., Marquenie, V. R., Pronk, J. T. & Daran, J.-M. G. Connecting central carbon and aromatic amino acid metabolisms to improve de novo 2-phenylethanol production in *Saccharomyces cerevisiae*. *Metab. Eng.* **56**, 165–180 (2019).
62. Gold, N. D. et al. Metabolic engineering of a tyrosine-overproducing yeast platform using targeted metabolomics. *Microb. Cell Factories* **14**, 73 (2015).
63. Sánchez, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
64. O'Brien, E. J. & Palsson, B. O. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Curr. Opin. Biotechnol.* **34**, 125–134 (2015).
65. Ye, C. et al. Comprehensive understanding of *Saccharomyces cerevisiae* phenotypes with whole-cell model WM\_S288C. *Biotechnol. Bioeng.* **117**, 1562–1574 (2020).
66. Ferreira, R. et al. Model-assisted fine-tuning of central carbon metabolism in yeast through dCas9-based regulation. *ACS Synth. Biol.* **2457–2463** (2019).
67. Aung, H. W., Henry, S. A. & Walker, L. P. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol.* **9**, 215–228 (2013).
68. Lewis, N. E. et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
69. Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
70. Jakočiūnas, T. et al. CasEMBLR: Cas9-facilitated multiloci genomic integration of *in vivo* assembled DNA parts in *Saccharomyces cerevisiae*. *ACS Synth. Biol.* **4**, 1226–1234 (2015).
71. Jensen, N. B. et al. EasyClone: method for iterative chromosomal integration of multiple genes in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **14**, 238–248 (2014).
72. Künzler, M., Paravicini, G., Egli, C. M., Irniger, S. & Braus, G. H. Cloning, primary structure and regulation of the *ARO4* gene, encoding the tyrosine-inhibited 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from *Saccharomyces cerevisiae*. *Gene* **113**, 67–74 (1992).
73. Helmstaedt, K., Strittmatter, A., Lipscomb, W. N. & Braus, G. H. Evolution of 3-deoxy-d-arabino-heptulosonate-7-phosphate synthase-encoding genes in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **102**, 9784–9789 (2005).
74. Luo, H. et al. Coupling S-adenosylmethionine-dependent methylation to growth: design and uses. *PLoS Biol.* **17**, e2007050 (2019).
75. Brooks, S., Gelman, A., Jones, G.L. & Meng, X.-L. *Handbook of Markov Chain Monte Carlo* (CRC, 2011).
76. Earl, D. J. & Deem, M. W. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7**, 3910–3916 (2005).
77. Rousseeuw, P. J. & Hubert, M. Robust statistics for outlier detection: robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**, 73–79 (2011).
78. Mockus, J. Application of Bayesian approach to numerical methods of global and stochastic optimization. *J. Glob. Optim.* **4**, 347–365 (1994).
79. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyperparameter optimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems* 2546–2554 (Curran Associates Inc., 2011).

## Acknowledgements

This work was supported by the Novo Nordisk Foundation and the European Commission Horizon 2020 programme (grant agreement no. 722287 and no. 686070). This work was also part of the DOE Agile BioFoundry (<http://agilebiofoundry.org>), supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, and the DOE Joint BioEnergy Institute (<http://www.jbei.org>), supported by the Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). H.G.M. was also supported by the Basque Government through the BERC 2014–2017 program, and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323. This work was also supported by the Chilean economic development agency, Corfo, through grant 17IEAT-73382.

## Author contributions

J.Z., S.D.P., J.D.K., J.N., and M.K.J. conceived the study. J.Z. and S.D.P. conducted all experimental work, Y.C. and B.J.S. all mechanistic modeling, and T.R., Z.C., and H.G.M. developed and applied statistical modeling and recommendations based on A.R.T., while E.A., A.R., A.P.-M., and M.J.F. developed and applied statistical modeling and recommendations based on TeselaGen EVOLVE model. S.D.P., J.Z., and M.K.J. wrote the manuscript.

## Competing interests

J.D.K. has a financial interest in Amyris, Lygos, Demetrix, Maple Bio, and Napigen. E.A., A.P.-M., A.R., and M.J.F. have a financial interest in TeselaGen Biotechnology. All other authors declare no competing interests.



**Additional information**

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41467-020-17910-1>.

**Correspondence** and requests for materials should be addressed to M.K.J.

**Peer review information** *Nature Communications* thanks Jean-Loup Faulon, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020